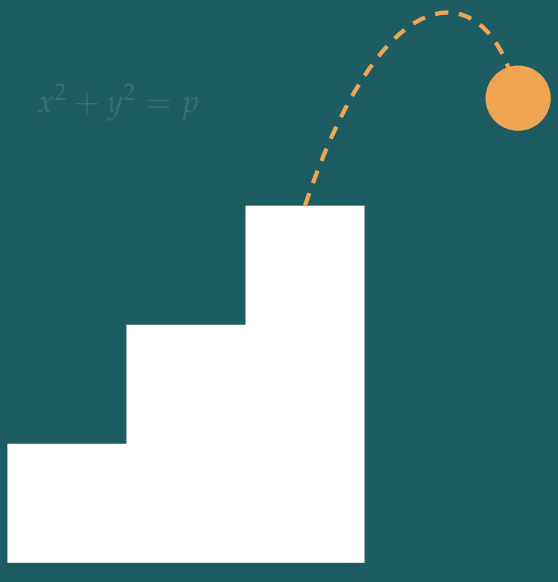
STEP 대비 · 대학 수학

Beyond STEP

제13장

세 상 속 의 통 계(statistics)와 무 작 위 성(randomness)



KEITH McNULTY

Beyond STEP

STEP 기출문제로 여는
대학 수학의 세계

제13장

세상 속의 통계(statistics)와 무작위성(randomness)

Keith McNulty

Beyond STEP: STEP 기출문제로 여는 대학 수학의 세계

제13장: 세상 속의 통계(statistics)와 무작위성(randomness)

이 PDF는 *Beyond STEP*의 전체 13개 장 가운데 제13장이다. 책 전체, 각 장의 개별 PDF, 모든 절의 인쇄용 연습지, 모든 STEP 문제의 자세한 풀이를 `beyondstep.keithmcnulty.org`에서 무료로 읽고 내려받을 수 있다.

2026년 11월 온라인 초판 발행.

이 PDF에서 지은이가 직접 만든 모든 것, 곧 본문, 풀이, 그림, 연습문제와 해답은 크리에이티브 커먼즈 CC0 1.0 Universal 라이선스에 따라 퍼블릭 도메인(public domain)으로 공개한다. 누구든 허락을 구하지 않고 어떤 목적으로든 복사하고, 고쳐 쓰고, 공유할 수 있다. 자세한 내용은 `creativecommons.org/publicdomain/zero/1.0`을 보라.

여기에서 다루는 STEP 문제의 저작권은 Cambridge University Press & Assessment에 있다. 문제는 이 책에 싣지 않았으며, 각 문제마다 공식 출처인 STEP Support Programme의 문제 데이터베이스나 OCR의 기출 시험지로 가는 링크를 달았다.

오류 제보는 `github.com/keithmcnulty/step_book_mcnulty`에 이슈로 남겨 주기 바란다.

장

13

세상 속의 통계와 무작위성

확률은 모형에서 데이터로 나아간다. 통계는 데이터에서 다시 세상으로 거슬러 올라간다.

이 장의 내용

13.1 신문 머리기사 읽기 STEP I 1992

13.2 들쥐 세기 STEP II 2000

13.3 독일 전차 문제 STEP III 1998

13.4 평균인가, 중앙값인가? STEP II 2000

13.5 곤란한 질문 던지기 STEP II 2023

13.6 쌓이는 오차 STEP II 1993

13.7 상관관계 만들어 내기 STEP III 2010

A 연습문제 해답 53쪽

데이터에서 배우기

> 모든 모형은 틀렸다. 그러나 그중 일부는 쓸모가 있다.
>
> — 조지 박스(George Box), 통계학자

확률(probability)에서는 주사위가 어떻게 조작되어 있는지 알고 앞으로 무슨 일이 일어날지를 묻는다. 통계학(statistics)에서는 무슨 일이 일어났는지를 보고 주사위가 어떻게 조작되어 있어야 하는지를 묻는다. 확률은 원인에서 결과로 추론하고, 통계학은 결과에서 원인으로 거슬러 올라가 추론한다. 개체 수에 대한 모든 추정값(estimate), 모든 여론조사, 모든 임상시험, 모든 과학적 측정은 통계적 추론(statistical inference)이다. 즉 제한되고 잡음이 섞인 데이터로부터 세상에 대해 이끌어 낸 결론이다.

통계학은 수학의 대부분의 분야에 비하면 젊은 학문이다. 오늘날의 모습은 20세기 초에 칼 피어슨(Karl Pearson), 로널드 피셔(Ronald Fisher), 예지 네이만(Jerzy Neyman) 등이 만들었으며, 그 동기는 흔히 실용적인 문제였다. 맥주 양조(t-검정은 기네스 양조장에서 발명되었다), 농작물 재배, 의약품 시험 같은 것들이다. 데이터가 폭발적으로 늘어난 오늘날, 통계학과 그 가까운 친척인 기계학습(machine learning)은 수학의 가장 중요한 응용 분야 가운데 하나이다.

통계학은 왜 중요한가

의약품, 공중보건, 경제 정책, 기후, 사법에 관한 결정은 모두 통계적 증거에 기반한다. 이 증거를 잘못 읽으면 실제로 대가를 치른다. 위양성(false positive) 결과에 겁을 먹은 환자(Section 12.2), 우연한 변동에 근거한 정책(Section 13.1), 작은 오차가 쌓여 가는 컴퓨터 시스템(Section 13.6)이 그 예이다. 데이터가 무엇을 말해 줄 수 있고 무엇을 말해 줄 수 없는지, 그리고 어떤 결론이든 얼마나 불확실한지를 이해하는 것은 깨어 있는 시민의 소양이며, 과학자, 의사, 공학자, 경제학자에게는 직업상 반드시 필요한 능력이다.

대학에서 만나게 될 내용

통계학 첫 강의에서는 추정(estimation: 알지 못하는 양의 좋은 추정값을 찾고 그것이 얼마나 정밀한지 밝히는 일), 가설 검정(hypothesis testing: 데이터가 어떤 가설과 부합하는지 판단하는 일), 신뢰구간(confidence interval), 회귀(regression: 한 변수를 다른 변수들과 관련짓는 일), 그리고 실험과 조사의 설계를 다룬다. 핵심 개념은 **표본분포**(sampling distribution: 추정값은 확률변수(random variable)이다), **편향**(bias)과 **분산**(variance), 그

리고 **가능도**(likelihood)이다. 이후의 강의에서는 베이즈 추론(Bayesian inference), 다변량 분석(multivariate analysis), 시계열(time series), 그리고 거대한 규모의 통계학이라 할 수 있는 기계학습의 방법들을 더 배운다.

이 장을 관통하는 질문들

세 가지 큰 질문

1. **실제 현상인가, 잡음인가?** 우연만으로 어느 정도의 변동을 예상해야 하며, 언제 변화를 진지하게 받아들여야 하는가?
2. **어떻게 추정해야 하는가?** 데이터가 주어졌을 때 알지 못하는 양의 가장 좋은 추정값은 무엇이며, 서로 다른 방법은 어떻게 비교하는가?
3. **무작위성은 어떻게 데이터에 들어오는가?** 데이터를 수집하고, 보호하고, 처리하는 방식은 우리의 결론에 어떤 영향을 미치는가?

이 장은 신문에 실린 통계를 비판적으로 읽는 것(Section 13.1)으로 시작하여 추정으로 넘어간다. 포획-재포획(capture–recapture)으로 들쥐의 수를 세고, 일련번호로부터 전체 개수를 추정하고, 평균(mean)과 중앙값(median)을 비교한다. 마지막으로 무작위성이 데이터에 들어오는 방식에 관한 세 가지 문제를 다룬다. 개인정보를 보호하는 무작위화된 조사, 쌓이거나 서로 상쇄되는 반올림 오차(rounding error), 그리고 정해진 상관(correlation)을 갖는 변수의 구성이다.

이 장의 아이디어

§	주제	아이디어
13.1	통계 읽기	비율, 자연 변동, 추세
13.2	포획-재포획	최대가능도(maximum likelihood)
13.3	일련번호	편향, 분산, 충분성(sufficiency)
13.4	중심 추정하기	효율성(efficiency)과 강건성(robustness)
13.5	무작위 응답(randomised response)	혼합(mixture)과 전체 분산(total variance)
13.6	반올림 오차	계통 오차(systematic error)와 우연 오차(random error)
13.7	상관	숄레스키(Cholesky) 구성

Section 13.7에서는 공분산(covariance)과 상관을 사용하는데, 이 둘은 2019년에 STEP 출제 범위(specification)에서 빠졌다. 필요한 내용은 그 절에서 설명한다.

이 장을 읽는 방법

모든 추정값은 확률변수라는 점을 기억하자. 데이터가 다르게 나왔다면 추정값도 달라졌을 것이다. 이 장의 각 방법에 대해, 여러 번 반복했을 때 그 추정값의 분포가 어떻게 될지 물어보자. 평균적으로 옳은가? 그리고 얼마나 변동하는가? 이 하나의 질문이 통계적 사고의 핵심이다.

13.1 신문 머리기사 읽기

문제

STEP I 1992, 14번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

`step.maths.org/questions/92-s1-q14`

도입

통계학(statistics)은 데이터로부터 배우는 학문이며, 그 첫 번째 교훈은 숫자가 스스로 말하지 않는다는 것이다. "여섯 배 더 안전하다", "놀라운 증가", "참사 증가": 날마다 신문 헤드라인은 숫자를 주장으로 바꾸고, 그 주장은 틀린 경우가 많다. 숫자가 거짓이어서가 아니라, 엉뚱한 대상과 비교하거나 평범한 무작위 변동을 추세로 착각하기 때문이다.

이 문제에는 공식이 전혀 없으며, 이는 STEP 문제로서는 드문 일이다. 이 문제는 통계학자처럼 생각하기를 요구한다. 각 헤드라인을 정당화하려면 무엇을 알아야 하는가? 그리고 순전히 우연만으로도 변동이 얼마나 클 것으로 예상되는가? 답에는 푸아송 분포(Poisson distribution)(Section 12.18)가 쓰이며, 여기서의 추론은 이 장의 나머지 부분의 토대가 된다.

시작하기 전에

생각해 보기 사망 사고가 일정한 비율로 서로 독립적이고 무작위로 일어난다면, 연간 총계는 얼마나 변동해야 하는가? 평균이 한 해에 1080이라면, 해마다 나타나는 전형적인 편차는 얼마인가?

서로 독립인 드문 사건의 개수는 근사적으로 푸아송 분포를 따르며, 평균이 μ인 푸아송 개수는 분산(variance)이 μ이고 표준편차(standard deviation)가 $\sqrt{\mu}$이다. $\mu = 1080$이면 표준편차는 약 33이다. 따라서 연간 총계가 약 1015에서 1145 사이라면 전혀 특별할 것이 없다.

차근차근 풀기

1단계 "걷는 것보다 비행기를 타는 것이 여섯 배 더 안전하다"

생각해 보기 보행자 사망자는 연간 1080명, 항공 사망자는 연간 180명이다. 정확히 무엇의 여섯 배인가?

이 헤드라인은 총 사망자 수를 비교하지만, 안전성은 **위험**(*risk*), 즉 노출(exposure) 단위당 사망자 수의 문제이다. 거의 모든 사람이 매일 걷는 반면, 비행기를 타는 사람은 비교적 적고 그마저도 가끔일 뿐이다. 공정한 비교를 하려면 이동한 킬로미터당, 이동 시간당, 또는 이동 1회당 사망자 수를 써야 하며, 이들은 매우 다른 답을 줄 수 있다(항공 여행은 킬로미터당으로는 극히 안전하지만, 이동 1회당으로는 그만큼 인상적이지 않다. 대부분의 항공 사망 사고가 이륙과 착륙 때 일어나기 때문이다). 주어진 수치로는 상대적 안전성에 대해 아무것도 "증명"할 수 없다. 또한 "도로 사고로 사망한 보행자"는 좁은 범주이며, 전체 도로 사망자 수는 훨씬 더 많다.

2단계 "하늘이 더 안전해졌다: 1990년 사망자 단 125명"

생각해 보기 항공 사망은 서로 독립인 드문 사건과 같은가? 연간 총계가 얼마나 변동할 것으로 예상하는가?

항공 사망은 무리 지어 일어난다. 추락 사고 한 번에 수백 명이 사망할 수 있다. 따라서 연간 총계는 주로 추락 사고의 횟수에 좌우되는데, 이 횟수는 작으며, 연간 총계는 해마다 극도로 크게 변동한다. 평균 180명에 비해 사망자가 125명인 해가 한 번 있었다는 것은 아무 변화도 없다는 것과 완전히 양립한다. 추락 사고가 한 번만 더 있었어도 수치는 완전히 달라졌을 것이다. 한 해의 데이터로는 추세를 보여 줄 수 없다.

3단계 "도로 참사 증가: 화요일에 7명 사망"

생각해 보기 평균적으로 하루에 보행자가 몇 명 사망하는가? 일일 사망자 수가 푸아송 분포를 따른다면, 하루에 7명 이상이 사망하는 경우를 얼마나 자주 보게 되는가?

평균은 하루에 보행자 사망자 $\frac{1080}{365} \approx 2.96$명이다. 일일 개수가 이 평균을 갖는 푸아송 분포를 따른다면 $P(\geq 7) \approx 0.031$이다. 즉 순전히 우연만으로도 1년에 약 11일은 보행자 사망자가 7명 이상일 것이다. (게다가 "도로 참사"에는 아마 운전자와 승객도 포함될 것이므로, 하루에 도로 사망자가 7명인 것은 더욱 놀랄 일이 아니다.) 하루의 데이터로는 추세에 대해 아무것도 말할 수 없다. 이 헤드라인은 변동이 매우 큰 계열에서 극단적인 값 하나를 골라낸 것으로, 전형적인 오류이다.

4단계 "놀라운 증가: 1990년 보행자 1350명 사망"

> **생각해 보기** 1350을 푸아송 표준편차와 비교하라. 이것은 단지 우연인가? 아니라면 달리 무엇으로 설명할 수 있으며, 그것이 "놀라운 증가"를 정당화하는가?

이것이 이 문제의 핵심 부분이다. 일정한 푸아송 모형 아래에서 1350은 평균 1080보다 270만큼 크며, 이는 약 8 표준편차에 해당한다. 우연일 가능성은 극히 낮다. 따라서 무언가가 변했을 가능성이 크다. 그러나 이 헤드라인은 여전히 여러 면에서 지나치다.

- **평균에 가려진 추세.** 수치 1080은 16년간의 평균이다. 사망자 수가 1974년 약 900명에서 1989년 약 1260명으로 꾸준히 증가했더라도 평균은 여전히 1080이었을 것이고, 1990년의 1350은 갑작스러운 "놀라운" 증가가 아니라 기존 추세의 연속일 뿐이다(그림 13.1.1). 연도별 수치가 필요하다.
- **노출.** 인구 증가, 보행 증가, 교통량 증가, 또는 인구 고령화는 모두 이동 1회당 도로가 더 위험해지지 않아도 사망자 수를 늘릴 것이다. 1인당 비율, 또는 걸은 킬로미터당 비율이 올바른 척도이다.
- **정의.** 사망을 기록하거나 분류하는 방식의 변화(예를 들어 사고 후 얼마 동안 일어난 사망까지 집계하는지)만으로도 수치가 급등할 수 있다.
- **과대산포**(*over-dispersion*). 실제 연간 개수는 날씨나 경제 활동 같은 조건이 변하기 때문에 푸아송 모형이 시사하는 것보다 더 크게 변동한다. "8 표준편차"는 어느 정도 신중하게 받아들여야 한다.

증가는 아마 실제이겠지만, 그것이 놀라운 일인지, 그리고 무엇이 원인인지 알려면 더 많은 데이터가 필요하다.

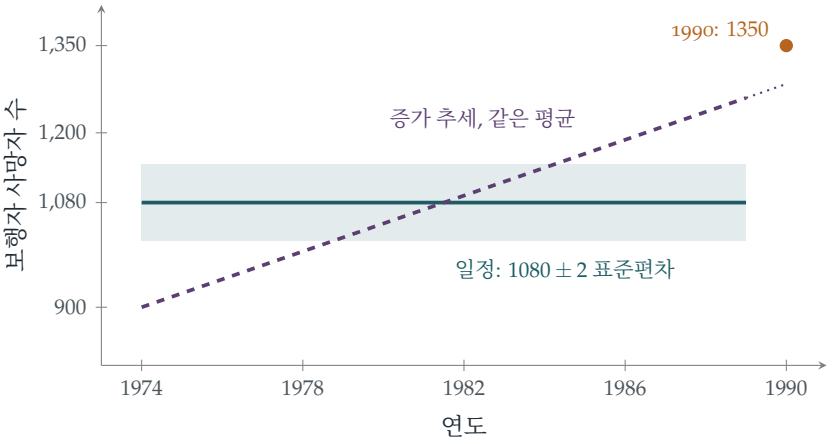


그림 13.1.1 16년 평균이 똑같이 1080인 매우 다른 두 역사. 사망자 수가 일정했다면(청록색, 전형적인 푸아송 변동의 띠와 함께), 1990년의 1350은 극적인 급등일 것이다. 꾸준히 증가해 왔다면(자주색), 1350은 추세보다 조금 높을 뿐이다. 평균만으로는 이 둘을 구별할 수 없다.

돌아보기

- **같은 것끼리 비교하라.** 위험에는 분모(denominator)가 필요하다. 1인당, 킬로미터당, 또는 이동 1회당 사망자 수이다.
- **자연스러운 변동을 알라.** 푸아송 변동성, 그리고 무리 지어 일어나는 사건의 훨씬 더 큰 변동성은 어떤 변화가 주목할 만한지 알려 준다.
- **평균은 추세를 가린다.** 요약 수치 하나로는 어떤 양이 시간에 따라 어떻게 변해 왔는지 설명할 수 없다.

숨은 수학

신호와 잡음

모든 통계적 질문은 근본적으로 **신호**(*signal*)(실제 효과나 추세)와 **잡음**(*noise*)(무작위 변동)을 분리하는 문제이다. 푸아송 모형은 서로 독립인 사건의 개수에 담긴 잡음을 재는 잣대, 곧 표준편차 $\sqrt{\mu}$를 준다. 드문 사건은 **상대적** 변동이 크다. 개수 9는 약 ± 3, 즉 그 크기의 3분의 1만큼 변동하지만, 개수 10 000은 약 1%만큼만 변동한다. 그래서 작은 수(하루의 사망자 수, 한 해의 추락 사고 수)가 그토록 자주 과잉 해석되며, 작은 마을의 암 "집단 발생"이 대개 우연으로 밝혀지는 것이다.

평균으로의 회귀

함정이 하나 더 있다. 도로 안전 카메라는 흔히 막 유난히 나쁜 한 해를 보낸 장소에 설치된다. 이듬해 사고가 줄어들면 카메라가 공을 인정받는다. 그러나 유난히 나쁜 해 다음에는 어차피 더 평범한 해가 올 가능성이 크다. 이것이 1880년대에 프랜시스 골턴(Francis Galton)이 처음 기술한 **평균으로의 회귀**(*regression to the mean*)이다. 카메라의 효과를 평균으로의 회귀와 분리하려면 카메라가 설치되지 않은 비슷한 장소와 비교해야 하며, 이것이 대조 실험(controlled experiment)의 바탕에 있는 생각이다.

플로렌스 나이팅게일의 통계

크림 전쟁의 간호사로 유명한 플로렌스 나이팅게일(Florence Nightingale)은 선구적인 통계학자이기도 했다. 1858년 그녀의 "장미 도표"는 부상보다 예방 가능한 질병으로 죽어 가는 병사가 훨씬 더 많다는 것을 보여 주었고, 정부가 군대의 위생을 개혁하도록 설득했다. 그녀는 원래의 총계가 아니라 비율(1년간 병사 천 명당 사망자 수)을 쓸 것과 같은 것끼리 비교할 것을 고집했는데, 바로 이 문제가 시험하는 점들이다. 그녀는 왕립통계학회(Royal Statistical Society)의 펠로로 선출된 최초의 여성이었다.

사고 기술

"무엇과 비교해서?"라고 물어라

통계적 주장을 받아들이기 전에, 무엇과 비교하는지, 분모가 무엇인지, 그리고 우연만으로 변동이 얼마나 클 것으로 예상되는지 물어라.

대학의 통계학 강좌에서는 이러한 질문을 정밀한 방법으로 바꾼다. 가설 검정(hypothesis test)은 관측값이 우연과 양립하는지를 묻고, 신뢰구간(confidence interval)은 어떤 양이 얼마나 정밀하게 측정되었는지 알려 주며, 실험 설계(design of experiments)는 비교가 공정하도록 보장한다.

더 도전하기

1 **푸아송 꼬리.** X가 평균이 2.96인 푸아송 분포를 따를 때 $\mathrm{P}(X \geq 7) \approx 0.031$임을 확인하라. 보행자 사망자가 9명 이상인 날은 1년에 며칠일 것으로 예상하는가?

2 **무리 지어 일어나는 사망.** 항공기 추락 사고가 연간 1.2회의 비율로 푸아송 과정(Poisson process)을 따라 일어나고, 사고마다 150명이 사망한다고 하자. 연간 항공 사망자 수의

평균과 표준편차를 구하라. 사망자가 125명 이하인 해는 이례적인가? 사망자가 정확히 125명인 해가 있을 수 있는가?

3 **유의한 증가?** 1974–1989년의 수치가 1080 ± 33으로 일정했다면, (정규 근사(normal approximation)를 이용하여) 한 해에 1350명 이상이 사망할 확률을 구하라.

4 **더 나은 헤드라인을 써라.** 네 헤드라인 각각에 대해, 데이터가 실제로 뒷받침하는 헤드라인을 써라.

더 읽을거리

- **David Spiegelhalter, *The Art of Statistics*** (Pelican). 뉴스에서 가져온 많은 예와 함께, 데이터로부터 배우는 방법을 다룬다.
- **Tim Harford, *How to Make the World Add Up*** (Bridge Street). 언론에 나오는 숫자에 대해 생각하는 열 가지 규칙.
- **Darrell Huff, *How to Lie with Statistics*** (Penguin). 1954년에 나온 짧은 고전으로, 지금도 여전히 유효하다.

13.2 들쥐 세기

문제

STEP II 2000, 13번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

`step.maths.org/questions/00-s2-q13`

도입

어떤 섬에 들쥐는 몇 마리나 살고 있을까? 들쥐를 하나하나 세는 것은 불가능하다. 숨고, 돌아다니고, 서로 비슷하게 생겼기 때문이다. 하지만 몇 마리를 잡아 표시를 한 뒤 놓아주고, 다시 몇 마리를 잡을 수는 있다. 200마리에 표시를 했고, 두 번째로 잡은 200마리 가운데 11마리에 표시가 되어 있다면, 전체 개체 수의 대략 $\frac{11}{200}$에 표시가 되어 있는 셈이다. 따라서 개체 수는 약 $200 \times \frac{200}{11} \approx 3636$이다.

이러한 **포획–재포획**(*capture–recapture*) 방법은 물고기, 새, 고래의 개체 수를 추정하는 데 쓰이며, 마약 사용자, 노숙인, 전쟁 사상자처럼 전부 셀 수 없는 모든 집단의 규모를 추정하는 데에도 쓰인다. 이 문제는 직관적인 추정을 통계적 추정에서 가장 중요한 아이디어인 **최대 가능도**(*maximum likelihood*) 원리를 이용해 탄탄한 기반 위에 올려놓는다. 그 원리란, 관측된 자료가 나올 확률을 가장 크게 만드는 값을 미지수의 값으로 고르라는 것이다.

시작하기 전에

생각해 보기 들쥐가 모두 N마리이고, 그 가운데 200마리에 표시가 되어 있다. 200마리를 임의로 표본으로 뽑는다. 정확히 11마리에 표시가 되어 있을 확률은 얼마인가?

같은 가능성을 가진 $\binom{N}{200}$가지 표본 가운데에서, 표시된 들쥐 200마리 중 11마리를, 표시되지 않은 $N-200$마리 중 189마리를 고르면 되므로

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

이것이 **초기하분포**(*hypergeometric distribution*)이다.

차근차근 풀기

1단계 가능도

N이 들어 있는 이항계수(binomial coefficient)를 풀어 쓰면

$$\binom{N-200}{189} = \frac{(N-200)!}{189!\,(N-389)!}, \qquad \binom{N}{200} = \frac{N!}{200!\,(N-200)!},$$

이므로

$$p_N = \binom{200}{11}\frac{200!}{189!}\cdot\frac{((N-200)!)^2}{N!\,(N-389)!} = k\frac{((N-200)!)^2}{N!\,(N-389)!}. \qquad \blacksquare$$

자료를 고정하고 미지수 N의 함수로 볼 때, p_N을 **가능도**(*likelihood*)라고 한다.

2단계 자연수 범위에서 최대화하기

> ? **생각해 보기** N은 자연수이므로 미분을 쓰기가 곤란하다. 대신 p_{N+1}과 p_N을 비교해 보자. 두 값의 비가 1보다 큰 것은 언제인가?

계승(factorial)은 대부분 약분된다.

$$\frac{p_{N+1}}{p_N} = \frac{(N-199)^2}{(N+1)(N-388)}.$$

이 값이 1보다 큰 것은 $(N-199)^2 > (N+1)(N-388)$일 때, 즉 $N^2 - 398N + 39601 > N^2 - 387N - 388$, 다시 말해 $11N < 39989$, $N < 3635.4$일 때이다. 따라서 p_N은 $N = 3636$까지 증가하고 그 뒤로 감소한다. 즉 최대가능도 추정값(maximum likelihood estimate)은

$$\hat{N} = 3636. \qquad \blacksquare$$

이것은 직관적인 추정값 $\frac{200\times 200}{11} = 3636.4$의 정수 부분이다.

3단계 세 번째 표본

두 번째 표본을 뽑은 뒤 그 안에 있던 표시되지 않은 들쥐 189마리에도 표시를 하므로, 이제 표시된 들쥐는 389마리이다. $N = 3636$으로 놓으면 표시되지 않은 들쥐는 $3636 - 389 = 3247$마리이다. 200마리로 이루어진 세 번째 표본에 표시된 들쥐가 정확히 j마리 들어 있을 확률은

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}.$$

$j = 0, 1, \ldots, 200$에 대한 이 확률들의 합은 1이어야 하므로, 문제의 항등식을 얻는다. ■

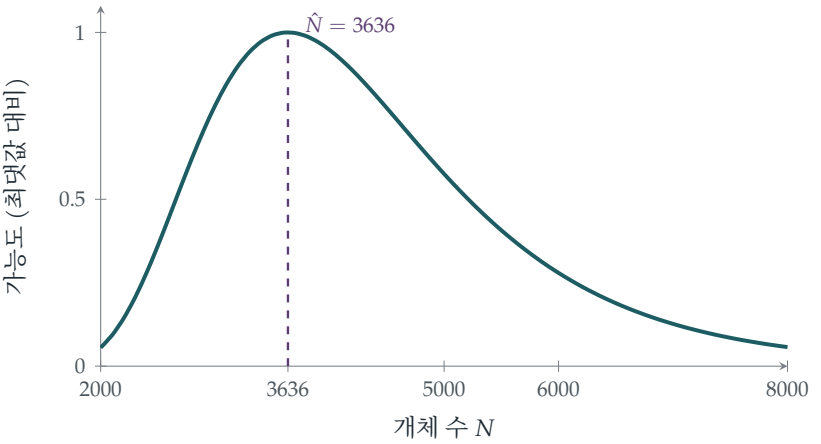


그림 13.2.1 두 번째 표본에서 표시된 들쥐가 11마리 나올 가능도 p_N을 개체 수 N의 함수로 나타낸 것으로, 최댓값이 1이 되도록 축척을 맞추었다. $N = 3636$에서 최대가 되지만, 폭이 넓고 한쪽으로 치우쳐 있다. 대략 2500에서 6500까지의 개체 수가 모두 자료와 상당히 잘 들어맞는다.

돌아보기

- **미지수의 함수로 본 자료의 확률.** 가능도는 보통의 관점을 뒤집는다. 자료는 고정되어 있고 모수(parameter)가 변한다.
- **이산적인 최대화에는 비를 이용한다.** 이웃한 항을 비교하여 정수를 변수로 하는 함수의 최댓값을 찾았다.
- **확률의 합은 1이다.** 모든 결과에 대해 분포를 더하여 조합론적 항등식(방데르몽드 항등식(Vandermonde's identity))을 증명했다.

숨은 수학

최대가능도

1910년대와 1920년대에 로널드 피셔(Ronald Fisher)가 발전시킨 **최대가능도** 원리는 다음과 같다. 미지의 모수를, 관측된 자료가 나올 확률을 가장 크게 만드는 값으로 추정하라. 이것은 단연 가장 널리 쓰이는 추정 방법으로, 자료에 곡선을 맞추는 일부터 기계 학습의 신경망을 훈련하는 일까지 두루 쓰인다. 신경망의 "손실 함수"는 대개 음의 로그 가능도이다. 피셔는 모형이 대부분의 일상적인 모형에서 성립하는 의미로 충분히 잘 정의되어 있다면, 표본이 클 때 최대가능도 추정값이 근사적으로 정규분포를 따르며 어떤 합리적인 추정량보다

도 뒤지지 않을 만큼 정밀하다는 것을 보였다. 중요한 예외도 있다. 다음 절(Section 13.3)에서는 미지수가 가능한 가장 큰 값, 즉 분포의 끝점인데, 이때 최대가능도 추정값은 사뭇 다르게 행동한다. 그 오차는 근사적으로 정규분포를 따르지 않으며, $\frac{1}{\sqrt{n}}$이 아니라 $\frac{1}{n}$처럼 줄어든다.

추정은 얼마나 정밀한가?

그림 13.2.1을 보면 가능도의 폭이 넓다. 11마리의 재포획은 그리 많은 정보가 아니다. 대략 계산해 보자. 두 번째 표본에 들어 있는 표시된 들쥐의 수는 평균이 $\frac{200\times 200}{N}$인 푸아송 분포(Poisson distribution)를 근사적으로 따르므로, 그 표준편차는 약 $\sqrt{11} \approx 3.3$이고, 상대오차는 약 30%이다. 이 오차는 N의 추정값에도 그대로 이어진다. N을 10% 이내로 추정하려면 생물학자들은 약 100마리를 재포획해야 한다. 통계학자들은 단순한 추정값의 작은 편향(bias)도 보정하는데, 이때 **채프먼 추정량**(*Chapman estimator*) $\frac{(M+1)(C+1)}{R+1} - 1$을 쓴다. 여기서 M과 C는 두 번 잡은 개체의 수이고, R은 재포획된 개체의 수이다.

라플라스, 프랑스 인구를 세다

라플라스(Laplace)는 1802년에 같은 아이디어를 이용하여 인구 조사 없이 프랑스의 인구를 추정했다. 그는 한 해 동안 나라 전체에 등록된 출생 수를 알고 있었고, 표본으로 고른 몇몇 코뮌(commune)에 대해서는 인구와 출생 수를 모두 알고 있었다. 인구와 출생 수의 비가 어디에서나 같다고 가정하여, 그는 프랑스의 인구를 약 28백만 명으로 추정했고, 불확실성의 척도까지 제시했다. 이 방법은 1890년대에 덴마크의 피오르에서 가자미에 표시를 한 C. G. J. Petersen과, 1930년대에 새 가락지를 이용한 Frederick Lincoln에 의해 동물 개체 수 추정에 다시 발견되었다. 그래서 지금도 흔히 **링컨–피터슨**(*Lincoln–Petersen*) 추정량이라고 부른다.

사고 기술

자료가 나올 가능성을 가장 크게 만드는 모수를 고르라

미지수를 추정하려면, 관측한 결과의 확률을 미지수의 함수로 쓰고 그것을 최대화하라.

대학에서 최대가능도는 통계적 추론의 뼈대이며, 추정값이 얼마나 정밀한지에 대한 이론(피셔 정보(Fisher information))과 경쟁하는 모형을 비교하는 방법(가능도비 검정(likelihood ratio test))이 그 위에 세워진다.

더 도전하기

1 **일반 공식.** M마리의 동물에 표시를 하고, C마리로 이루어진 두 번째 표본에 표시된 동물이 R마리 들어 있다면, N의 최대가능도 추정값이 $\frac{MC}{R}$의 정수 부분임을 보여라.

2 **방데르몽드.** 여자 a명과 남자 b명 가운데에서 c명으로 이루어진 위원회를 뽑는 방법의 수를 세어 $\sum_j \binom{a}{j}\binom{b}{c-j} = \binom{a+b}{c}$를 증명하여라.

3 **동전 던지기.** 앞면이 나올 확률 p를 모르는 동전을 n번 던져 앞면이 h번 나왔다. p의 최대가능도 추정값이 $\frac{h}{n}$임을 보여라.

4 **재포획이 없는 경우.** 두 번째 표본에 표시된 들쥐가 한 마리도 없다면 최대가능도 추정값은 어떻게 되는가? 이것은 이 방법에 대해 무엇을 말해 주는가?

더 읽을거리

- **David Spiegelhalter, *The Art of Statistics*** (Pelican). 추정과 불확실성에 관한 책이다.
- **George Seber, *The Estimation of Animal Abundance*** (Blackburn). 포획–재포획에 관한 표준 참고서이다.
- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). 최대가능도를 포함한 수리통계학 입문 교재이다.

13.3 독일 전차 문제

문제

STEP III 1998, 14번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

`step.maths.org/questions/98-s3-q14`

도입

제2차 세계 대전 중에 연합군 정보부는 독일이 전차를 몇 대나 생산하고 있는지 알고 싶어 했다. 첩보원들의 보고에 따르면 한 달에 약 1400대였다. 통계학자들은 다른 방법을 택했다. 노획한 전차의 변속기와 여러 부품에는 일련번호가 새겨져 있었는데, 이 일련번호는 사실상 $1, 2, \ldots, N$에서 뽑은 무작위 표본이다. 통계학자들은 이 번호만으로 생산량을 한 달에 약 250대로 추정하였다. 흔히 인용되는 수치에 따르면, 전쟁이 끝난 뒤 조사한 독일 측 기록에는 245대로 나와 있었다. 통계학자들이 첩보원들보다 훨씬 정확했던 것이다.

이 문제는 일련번호로부터 N을 추정하는 두 가지 방법을 비교한다. 두 방법 모두 평균적으로는 옳다. 그러나 한쪽이 다른 쪽보다 자료를 훨씬 효율적으로 사용하며, 어느 쪽이 그런지 알아내다 보면 추정 이론(estimation theory)의 핵심 아이디어에 이르게 된다. 평균적으로 옳은 추정량(estimator)들 가운데에서는 가장 적게 변동하는 것을 택하라는 것이다.

시작하기 전에

생각해 보기 M이 $[0, N]$ 위의 서로 독립인 균등분포(uniform distribution) 값 n개 중 가장 큰 값이라면, $\mathrm{P}(M \le m)$은 무엇인가?

$M \le m$일 필요충분조건은 모든 $X_i \le m$인 것이므로, $0 \le m \le N$에 대하여 $\mathrm{P}(M \le m) = \left(\frac{m}{N}\right)^n$이다(Section 12.25와 비교해 보라). 이를 미분하면 M의 확률밀도함수(probability density function)는 $\frac{nm^{n-1}}{N^n}$이다.

차근차근 풀기

1단계 두 추정량 모두 불편추정량이다

M에 대하여

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+1}N, \qquad \text{따라서} \qquad \mathrm{E}(Z_1) = \frac{n+1}{n}\cdot\frac{n}{n+1}N = N.$$

평균에 대하여: 각 X_i의 평균은 $\frac{N}{2}$이므로 $\mathrm{E}(Z_2) = \frac{2}{n}\cdot n \cdot \frac{N}{2} = N$이다. ■

두 추정량 모두 **불편추정량**(unbiased estimator)이다. 즉, 평균적으로는 올바른 답을 준다. 인수 $\frac{n+1}{n}$은 관측된 가장 큰 일련번호가 언제나 N보다 조금 작다는 사실을 보정해 준다.

2단계 두 추정량의 분산

> **생각해 보기** $\mathrm{E}(M^2)$을 구하고, 이를 이용하여 $\mathrm{Var}(M)$을 구하라. Z_2의 경우, $[0, N]$ 위의 균등분포를 따르는 확률변수 하나의 분산은 무엇인가?

$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+2}N^2$이므로

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{nN^2}{(n+1)^2(n+2)},$$
$$\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M) = \frac{N^2}{n(n+2)}.$$

$[0, N]$ 위의 균등분포를 따르는 확률변수의 분산(variance)은 $\frac{N^2}{12}$이고 X_i들은 서로 독립이므로

$$\mathrm{Var}(Z_2) = \frac{4}{n^2}\cdot n \cdot \frac{N^2}{12} = \frac{N^2}{3n}. \qquad ■$$

3단계 어느 쪽이 더 나은가?

> **생각해 보기** $\frac{1}{n(n+2)}$과 $\frac{1}{3n}$을 비교하라. n이 커질 때 각각 어떻게 변하는가?

두 분산의 비는 $\frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_2)} = \frac{3}{n+2}$이며, 이는 $n \geq 2$일 때 1보다 작다($n = 1$일 때는 두 추정량이 일치하므로 비가 1이다). 따라서 Z_1이 더 낫고, n이 클 때는 훨씬 더 낫다. Z_1의 표준편차(standard deviation)는 $\frac{1}{n}$처럼 줄어들지만, Z_2의 표준편차는 $\frac{1}{\sqrt{n}}$처럼 줄어들 뿐이다(그림 13.3.1). 일련번호가 $n = 10$개이면 Z_1의 표준편차는 N의 약 9%이고, Z_2는 약 18%이다.

Z_1을 택할 이유가 하나 더 있다. Z_2는 터무니없는 값을 줄 수 있다. 일련번호가 10, 20, 300이면 $Z_2 = \frac{2}{3}\cdot 330 = 220$인데, 이는 실제로 관측한 일련번호보다도 작다. 반면 $Z_1 = \frac{4}{3}\cdot 300 = 400$은 결코 가장 큰 관측값보다 작아질 수 없다. ■

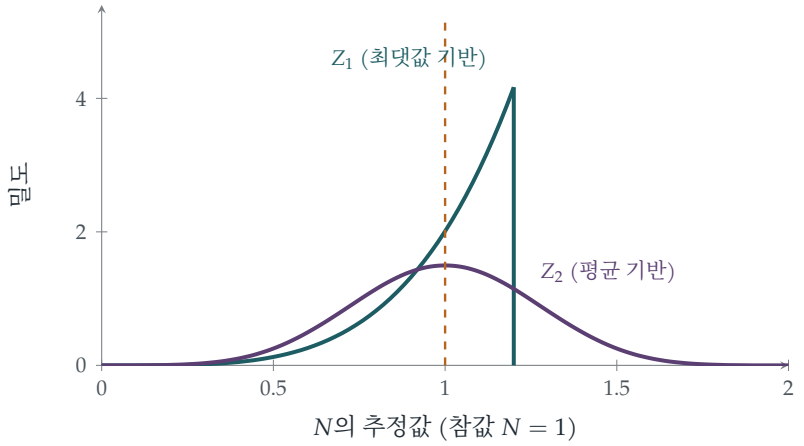


그림 13.3.1 일련번호가 $n = 5$개이고 참값이 $N = 1$일 때 두 추정량의 표본분포(sampling distribution). 둘 다 평균은 정확히 1이지만, Z_1(청록색)은 Z_2(자주색)보다 참값 근처에 훨씬 더 촘촘히 모여 있으며, 관측된 가장 큰 일련번호보다 작아지는 일이 결코 없다.

돌아보기

- **불편성.** 각 추정량은 기댓값이 정확히 N이 되도록 크기를 조정한 것이다.
- **분산을 비교하라.** 불편추정량들 가운데에서는 분산이 더 작은 것이 더 믿을 만하다.
- **가장 많은 정보를 담은 통계량을 사용하라.** 최댓값은 표본이 지닌 N에 관한 정보를 모두 담고 있지만, 평균은 그 정보의 상당 부분을 버린다.

숨은 수학

충분성

최댓값이 왜 이렇게 훨씬 나은가? 가장 큰 값 M이 주어지면, 나머지 값들은 N이 무엇이든 그저 $[0, M]$ 위의 균등분포를 따를 뿐이다. 즉, 나머지 값들은 N에 관해 더 이상 아무런 정보도 주지 않는다. 최댓값은 **충분통계량**(sufficient statistic)이다. C. R. Rao와 David Blackwell의 정리(1945–47)에 따르면, 어떤 불편추정량이든 충분통계량이 주어졌을 때의 조건부 기댓값으로 바꾸면 개선되거나, 적어도 나빠지지는 않는다. 여기서는 그렇게 하면 Z_2가 Z_1이 된다. 나아가 **레만–셰페 정리**(Lehmann–Scheffé theorem)를 쓰면 Z_1이 가능한 가장 작은 분산을 갖는 불편추정량임을 보일 수 있다. 또한 Z_1의 분산은 $\frac{1}{n^2}$처럼 줄어드는데, 이는 대부분의 추정량이 보이는 $\frac{1}{n}$보다 빠르다는 점에도 주목하라. 이런 일이 생기는 까닭은 N이 분포의 끝점이고, 자료가 그 끝점 쪽으로 몰려 쌓이기 때문이다.

이산적인 경우

실제 일련번호는 자연수 $1, 2, \dots, N$이며, 비복원추출(sampling without replacement)로 뽑힌다. 이때 가장 좋은 불편추정값은 $\hat{N} = m + \frac{m}{k} - 1$인데, 여기서 m은 관측한 일련번호 k개 중 가장 큰 값이다. 즉, 최댓값에 관측값 사이의 평균 간격을 더한 것이다. 주로 변속기와 바퀴의 일련번호에 바탕을 둔 이런 종류의 추정값을 써서, 연합군 통계학자들은 독일의 생산량을 한 달에 약 246대로 추정하였다. 나중에 독일 측 기록에서 확인된 수치는 245대였고, 첩보 보고가 제시한 수치는 1400대였다.

역사 속의 일련번호

이 방법은 그 뒤로도 여러 번 쓰였다. 예를 들어 분석가들은 처음 판매된 몇 개의 일련번호로부터 새로운 소비자 제품이 몇 개나 만들어졌는지를 추정하였다. 제조사들도 교훈을 얻어, 지금은 많은 제조사가 일련번호를 단순한 순서로 읽히지 않도록 암호화한다. Richard Ruggles와 Henry Brodie는 1947년 *Journal of the American Statistical Association*에 전시의 이 연구를 소개하였다.

사고 기술

추정량은 표본분포로 판단하라

추정량은 확률변수이다. 평균적으로 옳은지(편향)와 얼마나 변동하는지(분산)를 묻고, 오차가 더 작은 쪽을 택하라.

대학에서는 이것이 효율적 추정(efficient estimation) 이론으로 이어진다. 크라메르–라오 하한(Cramér–Rao bound)은 불편추정량의 분산이 얼마나 작아질 수 있는지를 알려 주고, 충분성은 그 하한에 어떻게 도달하는지를 설명해 준다.

더 도전하기

1. **평균제곱오차.** 편향추정량 M에 대하여 $\mathrm{E}(M - N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M) - N\big)^2$이다. 이 평균제곱오차(mean squared error)를 계산하고 $\mathrm{Var}(Z_1)$과 비교하라. $\mathrm{E}(cM - N)^2$을 최소로 하는 상수 c를 구하라.

2. **최솟값도.** $(n+1)\min\{X_i\}$도 N의 불편추정량임을 보이고, 그 분산을 구하라. 이 추정량은 쓸 만한가?

3 **이산 공식.** 일련번호 $1, \dots, N$에서 k개를 비복원추출하고, 그 최댓값을 m이라 하자. $\mathrm{P}(\max \le m) = \binom{m}{k}/\binom{N}{k}$임을 보이고, $k = 1$일 때 $m + \frac{m}{k} - 1$이 불편추정량임을 확인하라.

4 **전차.** 노획한 전차 5대의 일련번호가 19, 40, 42, 60, 98이라면, 만들어진 전차의 수를 추정하라.

더 읽을거리

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**, *Journal of the American Statistical Association* 42 (1947). 원래의 기록이다.
- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). 편향, 분산, 충분성을 다룬다.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican).

13.4 평균인가, 중앙값인가?

문제

STEP II 2000, 14번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

`step.maths.org/questions/00-s2-q14`

도입

측정값들의 중심을 추정할 때 평균(mean)을 쓸 수도 있고 중앙값(median)을 쓸 수도 있다. 어느 쪽이 더 나을까? 답은 데이터에 따라 달라진다. 가끔 터무니없는 오차가 섞이는 측정값이라면 중앙값이 훨씬 믿을 만하다. 엉뚱한 값 하나가 평균은 어디로든 끌고 갈 수 있지만 중앙값은 거의 움직이지 못한다. 반면 얌전한 데이터라면 대개 평균이 낫다. 이 문제는 균등분포(uniform distribution)를 따르는 데이터에서 평균이 얼마나 더 나은지를 측정한다. 중앙값의 분산(variance)은 평균의 분산의 세 배이므로, 중앙값이 평균만큼 정밀해지려면 관측값이 세 배 필요하다.

이 비교를 통해 통계적 방법을 고를 때의 주요 개념 가운데 하나인 **효율**(*efficiency*)과, Section 4.4에서도 나왔던 베타 적분(beta integral)을 소개한다.

시작하기 전에

? **생각해 보기** 중앙값의 확률밀도함수(density)가 왜 $y^n(1-y)^n$에 비례할까? (Y가 y 근처에 있으려면 $2n+1$개의 값 가운데 몇 개가 y보다 작아야 하고, 몇 개가 커야 할까?)

중앙값이 y에 있으려면 값 하나는 y에 있고, n개는 그보다 작고(각각 확률 y), n개는 그보다 커야 한다(각각 확률 $1-y$). 이로부터 인수 $y^n(1-y)^n$이 나온다. 상수는 어느 값이 어디에 속하는지를 고르는 방법의 수를 센 것이다.

차근차근 풀기

1단계 상수

밀도함수를 적분하면 1이 되어야 한다.

$$k\int_0^1 y^n(1-y)^n\,\mathrm{d}y = k\cdot\frac{n!\,n!}{(2n+1)!} = 1, \qquad k = \frac{(2n+1)!}{(n!)^2}. \qquad \blacksquare$$

2단계 평균과 분산

밀도함수는 $\frac{1}{2}$에 대하여 대칭(symmetric)이므로(y를 $1-y$로 바꾸어 보라) $\mathrm{E}(Y)=\frac{1}{2}$이다.

> **생각해 보기** 베타 적분을 이용하여 $\mathrm{E}(Y^2) = k\int_0^1 y^{n+2}(1-y)^n\,\mathrm{d}y$를 계산하고 계승을 정리하라.

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

따라서

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}. \qquad \blacksquare$$

3단계 평균과 비교하기

> **생각해 보기** 서로 독립인 균등분포 확률변수 $2n+1$개의 평균 $\bar{X}$의 평균과 분산은 무엇인가?

각 X_i의 평균은 $\frac{1}{2}$, 분산은 $\frac{1}{12}$이므로 $\mathrm{E}(\bar{X})=\frac{1}{2}$이고 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)}$이다. n이 클 때

$$\mathrm{Var}(Y) \approx \frac{1}{8n}, \qquad \mathrm{Var}(\bar{X}) \approx \frac{1}{24n}.$$

> **생각해 보기** Y와 $\bar{X}$가 평균이 $\frac{1}{2}$이고 표준편차(standard deviation)가 각각 σ_Y, $\sigma_{\bar{X}}$인 정규분포를 따른다면, 두 확률을 표준정규분포함수(standard normal distribution function) Φ로 나타내라.

$\mathrm{P}\left(|Y-\frac{1}{2}| < \frac{d}{\sqrt{n}}\right) = 2\Phi\left(\frac{d}{\sqrt{n}\,\sigma_Y}\right) - 1$이고 $\sqrt{n}\,\sigma_Y \to \frac{1}{\sqrt{8}}$이다. $\bar{X}$도 마찬가지로 $\sqrt{n}\,\sigma_{\bar{X}} \to \frac{1}{\sqrt{24}}$이다. 따라서 두 확률은 각각 $2\Phi(d\sqrt{8})-1$과 $2\Phi(d\sqrt{24})-1$에 수렴한다. Φ는 증가함수이고 $\sqrt{24} > \sqrt{8}$이므로 두 번째 값이 더 크고, 따라서 충분히 큰 모든 n에 대하여 부등식이 성립한다. ■평균이 참 중심에 가까이 있을 가능성이 더 크다.

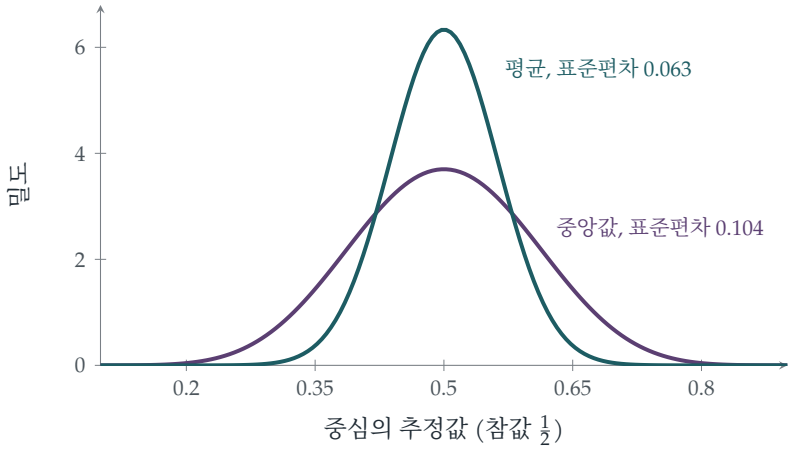


그림 13.4.1 서로 독립인 균등분포 관측값 21개의 중앙값(자주색)과 평균(청록색)의 표본분포. 둘 다 참값 $\frac{1}{2}$를 중심으로 하지만 평균이 더 촘촘하게 모여 있다. 평균의 분산은 중앙값의 분산의 약 3분의 1이다.

돌아보기

- **베타 적분.** $\int_0^1 y^r(1-y)^s\,\mathrm{d}y$ 하나의 공식으로 정규화 상수(normalising constant)와 적률(moments)을 모두 구했다.
- **대칭성으로 평균을 구했다.** $\mathrm{E}(Y)$를 구하는 데 적분이 필요 없었다.
- **퍼짐 정도는 알맞은 척도에서 비교한다.** 표준편차에 $\sqrt{n}$을 곱하자 비교가 n에 의존하지 않게 되었다.

숨은 수학

효율

두 불편추정량(unbiased estimator)의 **상대 효율**(*relative efficiency*)은 두 분산의 비이다. 이는 한 추정량이 다른 추정량과 같은 정밀도를 얻으려면 관측값이 몇 배 필요한지를 말해 준다. 균등분포 데이터에서 중앙값의 평균에 대한 효율은 $\frac{1}{3}$이다. 정규분포 데이터에서는 $\frac{2}{\pi} \approx 0.64$로, 중앙값은 관측값이 약 57% 더 필요하다. 그러나 꼬리가 두꺼운(heavy-tailed) 데이터에서는 상황이 뒤바뀐다. 코시 분포(Cauchy distribution, Section 12.22)에서는 관측값 n개의 평균이 관측값 하나보다 나을 것이 없으므로 평균은 쓸모가 없지만, 중앙값은 잘 작동한다. 라플라스 분포(Laplace distribution, Section 12.24)에서는 중앙값이 평균보다 두 배 효율적이다.

로버스트 통계

중앙값은 **로버스트**(*robust*)하다. 데이터의 거의 절반을 임의의 값으로 바꾸어도 중앙값은 크게 움직이지 않는다. 평균에는 이런 보호 장치가 없어서 잘못된 값 하나만으로도 망가질 수 있다. 1960년대와 1970년대에 존 튜키(John Tukey), Peter Huber 등은 **로버스트 통계**(*robust statistics*)를 발전시켰다. 이는 깨끗한 데이터에서는 평균에 거의 버금가는 효율을 보이면서도 이상값(outlier)에 잘 견디는 추정량들로, 예를 들어 **절사평균**(*trimmed mean*)(가장 큰 값과 가장 작은 값 10%씩을 버리고 나머지를 평균한 것)이 있다. 절사평균은 다이빙이나 체조 같은 스포츠의 채점에 쓰인다.

라플라스와 중앙값

라플라스(Laplace)는 1774년에 오차가 이중지수분포(라플라스 분포)를 따르면 참값의 가장 좋은 추정값은 중앙값임을 보였다. 가우스(Gauss)는 1809년에 오차가 정규분포를 따르면 가장 좋은 추정값은 평균임을 보였고, 이로부터 최소제곱법(method of least squares)에 이르렀다. 어느 쪽이 옳은지는 오차가 어떻게 행동하는지에 달려 있으며, 통계학자들은 새로운 종류의 데이터가 나올 때마다 여전히 이 문제를 논쟁한다.

사고 기술

방법을 변동성으로 비교하기

두 방법이 모두 평균적으로는 옳을 수 있다. 그 답이 얼마나 흔들리는지, 그리고 데이터가 많아질수록 그 흔들림이 얼마나 빨리 줄어드는지를 비교하라.

대학에서는 큰 표본에서 분산을 비교하는 점근적 효율(asymptotic efficiency)이, 모형이 틀렸을 수 있는 경우를 위한 로버스트성과 함께, 통계적 방법을 고르는 표준 도구로 쓰인다.

더 도전하기

1 **중앙값의 밀도함수.** 값 하나는 $[y, y+\delta y]$에, n개는 그 아래에, n개는 그 위에 있는 경우를 세어 밀도함수 $\frac{(2n+1)!}{(n!)^2} y^n (1-y)^n$을 유도하라.

2 **관측값 세 개.** $n = 1$(균등분포 관측값 세 개)일 때 $\mathrm{Var}(Y)$와 $\mathrm{Var}(\bar{X})$를 정확히 구하라. 어느 쪽이 더 작은가?

3 **중간범위.** 중심의 또 다른 추정량으로 중간범위(midrange) $\frac{1}{2}(\max + \min)$가 있다. 관측값이 $m = 2n + 1$개인 균등분포 데이터에서 그 분산이 $\frac{1}{2(m+1)(m+2)}$임을 보여라. 평균과 비교하면 어떠한가? (Section 13.3와 비교해 보라.)

4 **이상값.** 측정값 열한 개가 9.8, 9.9, 10.0, 10.0, 10.1, 10.1, 10.2, 10.2, 10.3, 10.4, 104(입력 실수)이다. 평균과 중앙값을 비교하라.

더 읽을거리

- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). 순서통계량(order statistics)과 효율을 다룬다.
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods*** (Springer). 로버스트 방법이 왜 중요한지를 비전문가에게 설명한다.
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom*** (Harvard University Press). 자료의 집계, 평균과 중앙값을 역사적 관점에서 살펴본다.

13.5 곤란한 질문 던지기

문제

STEP II 2023, 11번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2023년 STEP 시험지 묶음(zip 파일) 가운데 STEP II 시험지에 있다:

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

도입

시험에서 부정행위를 한 학생이 몇 분의 몇인지, 혹은 직장에서 물건을 훔친 직원이 몇 분의 몇인지 어떻게 알아낼 수 있을까? 직접 물어보면 많은 사람이 거짓말을 할 것이다. 1965년 통계학자 Stanley Warner는 기발한 해결책인 **확률화 응답**(*randomised response*)을 고안했다. 각 사람은 몰래 동전을 던지거나 주사위를 굴리고, 그 결과에 따라 민감한 질문에 사실대로 답할지, 아니면 그 반대 질문에 답할지가 정해진다. 면접자는 어느 쪽인지 결코 알 수 없으므로 어떤 개인의 답도 그 사람에게 불리한 증거가 되지 않는다. 그런데도 확률화에 쓰인 확률을 알고 있으므로 전체 비율은 여전히 추정할 수 있다.

이 문제는 그 뒤에 숨은 수학을 탐구한다. 각 사람을 **독립적으로**(*independently*) 확률화하면(소문항 (ii)) 비율을 정확히 짚어 내는 잘 행동하는 합계를 얻는다. 반면 모든 사람을 **단 한 번의** 동전 던지기로 확률화하면(소문항 (iii)) 평균은 같지만 봉우리가 두 개인 전혀 다른 분포의 합계가 나오고, 이것은 우리에게 거의 아무것도 알려 주지 않는다. 무작위성을 어디에 두느냐가 중요하다.

시작하기 전에

생각해 보기 Y는 확률 p로 X_1이고 확률 q로 X_2이다. $\mathrm{E}(Y)$는 무엇인가? $\mathrm{E}(Y^2)$는 무엇인가?

어느 변수가 선택되는지에 대해 조건화(conditioning)하면 $\mathrm{E}(Y) = p\mu_1 + q\mu_2$이고, $\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2)$이다.

차근차근 풀기

1단계 혼합분포의 분산

$$\mathrm{Var}(Y) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2) - (p\mu_1 + q\mu_2)^2.$$

$p - p^2 = pq$와 $q - q^2 = pq$를 이용하면 μ가 들어간 항은 $p\mu_1^2 + q\mu_2^2 - p^2\mu_1^2 - 2pq\mu_1\mu_2 - q^2\mu_2^2 = pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2$이다. 따라서

$$\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2. \qquad \blacksquare$$

혼합분포(mixture)의 분산(variance)은 각 성분 내부 분산의 평균에 성분들의 평균 사이의 분산을 더한 것이다.

2단계 독립적인 확률화

> **생각해 보기** $\mathrm{P}(B = 1)$은 무엇인가? B의 n개 값은 서로 독립인가?

$\mathrm{P}(B = 1) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{5}{6} = \frac{1}{2}$이다. n개의 값은 서로 독립이고 각각의 성공 확률이 $\frac{1}{2}$이므로 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$이며, 평균은 $\frac{n}{2}$, 표준편차(standard deviation)는 $\frac{\sqrt{n}}{2}$이다.

평균의 10% 이내라는 것은 $|Z_1 - \frac{n}{2}| < \frac{n}{20}$이라는 뜻이다. 표준화한 단위로는 $|Z| < \frac{n/20}{\sqrt{n}/2} = \frac{\sqrt{n}}{10}$이고, 이 값은 무한대로 발산하므로 확률은 1로 수렴한다. $\blacksquare$

3단계 모두를 위한 동전 하나

> **생각해 보기** 이제 동전 던지기 한 번이 n개의 주사위 전부에 대해 결정한다. 앞면이 나왔을 때 Z_2의 분포는 무엇인가? 뒷면이 나왔을 때는? 소문항 (i)을 이용하라.

앞면이 나오면 $Z_2 \sim \mathrm{B}(n, \frac{1}{6})$이고, 평균은 $\frac{n}{6}$, 분산은 $\frac{5n}{36}$이다. 뒷면이 나오면 $Z_2 \sim \mathrm{B}(n, \frac{5}{6})$이고, 평균은 $\frac{5n}{6}$, 분산은 $\frac{5n}{36}$이다. $p = q = \frac{1}{2}$로 소문항 (i)을 적용하면

$$\mathrm{E}(Z_2) = \frac{n}{2}, \qquad \mathrm{Var}(Z_2) = \frac{5n}{36} + \frac{1}{4}\left(\frac{4n}{6}\right)^2 = \frac{5n}{36} + \frac{n^2}{9}.$$

> **생각해 보기** 확률의 대부분은 어디에 있는가? 평균 $\frac{n}{2}$ 근처에 확률이 많이 있는가?

Z_2는 $\frac{n}{6}$과 $\frac{5n}{6}$ 근처에 폭이 대략 $\sqrt{n}$인 봉우리 두 개를 가지며(그림 13.5.1), 그 사이에는 거의 아무것도 없다. 평균과 분산이 같은 정규분포(normal distribution)라면 $\frac{n}{2}$를 중심으로 하는 봉우리 하나일 텐데, 그곳은 바로 Z_2가 거의 놓이지 않는 자리이다. Z_2가 $\frac{n}{2}$의 10% 이내에

있으려면 $\mathrm{B}(n, \frac{1}{6})$ 또는 $\mathrm{B}(n, \frac{5}{6})$ 확률변수가 $[0.45n, 0.55n]$에 놓여야 한다. $\mathrm{B}(n, \frac{1}{6})$의 경우 이는 평균보다 적어도 $0.28n$만큼 위에 있다는 뜻이고, 이는 표준편차로 약 $\frac{0.28n}{0.37\sqrt{n}} \approx 0.75\sqrt{n}$배에 해당하므로 확률은 0으로 수렴한다. $\mathrm{B}(n, \frac{5}{6})$도 마찬가지이다. ■

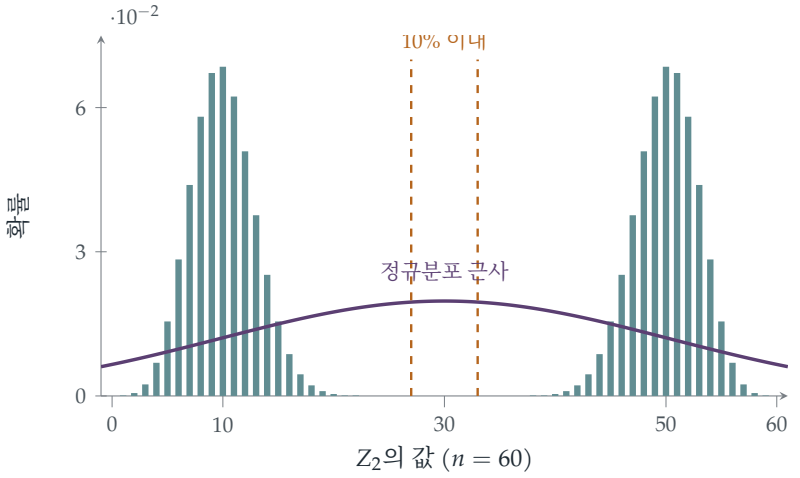


그림 13.5.1 주사위 $n = 60$개에 대한 Z_2의 분포(막대): 한 번의 동전 던지기에서 나올 수 있는 두 결과 때문에 10과 50에 봉우리가 두 개 생긴다. 평균(30)과 분산이 같은 정규분포(자주색)는 잘 맞지 않으며, 평균의 10% 이내(호박색 선 사이)에는 확률이 거의 없다.

돌아보기

- **숨겨진 선택에 대해 조건화하라.** 혼합분포의 평균과 분산은 어느 성분이 선택되었는지에 대해 조건화하여 얻었다.
- **내부 분산 더하기 사이 분산.** 추가된 항 $pq(\mu_1 - \mu_2)^2$은 두 성분 사이의 퍼짐을 측정했다.
- **평균과 분산이 전부는 아니다.** Z_2는 Z_1과 평균이 같았지만 모양은 전혀 달랐다.

숨은 수학

확률화 응답

Warner의 방식에서 각 응답자는 남몰래 선택하여, 확률 p로 "당신은 집단 A에 속합니까?"에 답하고, 확률 $1 - p$로 "당신은 집단 A에 속하지 **않습니까**?"에 답한다. 비율 π만큼의 사람이 A에 속한다면 "예"라고 답할 확률은

$$\lambda = p\pi + (1 - p)(1 - \pi),$$

이므로, 관측된 "예"의 비율 $\hat{\lambda}$로부터 $\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$로 추정할 수 있다(단, $p \neq \frac{1}{2}$). 문제의 소문항 (ii)가 바로 이것이며, 각자의 주사위가 결정을 내린다. 프라이버시의 대가는 정밀도이다. $2p-1$로 나누면 분산이 커지므로, 확률화된 조사는 직접 조사보다 더 많은 응답자가 필요하다. 하지만 각 사람이 독립적으로 확률화하므로 큰수의 법칙(law of large numbers)이 작동한다. 소문항 (iii)은 확률화를 공유하면 무엇이 잘못되는지 보여 준다. 단 하나의 동전에 대한 불확실성은 결코 평균화되어 사라지지 않는다.

차분 프라이버시

확률화 응답은 **차분 프라이버시**(*differential privacy*)의 조상이다. 차분 프라이버시는 2006년 Cynthia Dwork와 동료들이 도입한 프라이버시의 수학적 정의로, 어떤 한 사람의 데이터를 추가하거나 제거해도 모든 출력의 확률이 기껏해야 작은 배수만큼만 변한다면 그 절차는 프라이버시를 보장한다고 한다. 기술 기업들은 이를 이용해 사용 통계를 수집하며, 미국 인구조사국(United States Census Bureau)은 2020년 인구조사에서 이를 사용하여 공개되는 수치에 신중하게 보정한 무작위 잡음을 더함으로써 어떤 개인의 응답도 재구성할 수 없게 했다.

전분산 법칙

소문항 (i)의 항등식은 **전분산 법칙**(*law of total variance*)의 특수한 경우이다. 즉 $\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid C)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid C)\big)$이며, 여기서 C는 숨겨진 선택이다. 이 법칙은 **분산분석**(*analysis of variance*)(ANOVA)의 바탕이 된다. 분산분석은 1920년대에 피셔(Ronald Fisher)가 로섬스테드(Rothamsted)에서의 농업 실험을 위해 도입한 것으로, 작물 수확량의 변동을 비료나 품종으로 설명되는 부분과 설명되지 않고 남는 부분으로 나눈다.

사고 기술

요약값만 보지 말고 분포 전체를 보라

평균과 분산은 봉우리가 하나인 분포는 잘 설명하지만, 혼합분포에 대해서는 심하게 오해를 불러일으킬 수 있다. 확률이 실제로 어디에 있는지 물어라.

대학에서는 여러 집단으로 이루어진 모집단을 설명하는 데 혼합 모형(mixture model)을 사용하며, 전분산 법칙은 조사 설계에서 계층적 모형(hierarchical model)에 이르기까지 통계학의 표준 도구이다.

더 도전하기

1 **쌍둥이.** STEP II 2005, 12번 문제 (`step.maths.org/questions/05-s2-q12`)를 풀어 보라. 이것은 겉모습만 바꾼 확률화 응답 퍼즐이다. 어떤 답이 동전에 대해 아무것도 알려 주지 않는 것은 언제인가?

2 **π 추정하기.** $p = \frac{3}{4}$인 Warner 조사에서 400명이 응답했고 160명이 "예"라고 답했다. 집단 A의 비율 π를 추정하고, 그 추정값의 근사적인 표준오차(standard error)를 구하라.

3 **전분산.** 여러 값을 갖는 이산 숨은 변수 C에 대해 전분산 법칙을 증명하라.

4 **공유된 동전.** 소문항 (iii)에서 $Z_2 > \frac{n}{2}$일 확률은 얼마인가? Z_2가 $\frac{n}{2}$에 가까운 경우는 거의 없는데도, n이 클 때 이 확률이 $\frac{1}{2}$에 가까운 이유를 설명하라.

더 읽을거리

- **Stanley Warner, "Randomized response: a survey technique for eliminating evasive answer bias"**, *Journal of the American Statistical Association* 60 (1965).
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*** (온라인 무료 공개). 현대 이론을 다루며, 앞부분의 장들은 쉽게 읽을 수 있다.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican). 조사와 그 함정을 다룬다.

13.6 쌓이는 오차

문제

STEP II 1993, 15번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

`step.maths.org/questions/93-s2-q15`

도입

1982년 밴쿠버 증권거래소는 1000에서 출발하는 새로운 지수를 도입했다. 주가가 바뀔 때마다 지수를 다시 계산하고 소수점 아래 셋째 자리까지만 남기고 **버림**(truncation)했다. 한 번 버릴 때마다 평균적으로 0.0005를 잃었는데, 이는 무시해도 될 만큼 작아 보인다. 그러나 재계산은 하루에 약 3000번 이루어졌고, 22개월이 지나자 시장은 올랐는데도 지수는 520 근처에 머물러 있었다. 지수를 올바르게 다시 계산하자 하룻밤 사이에 약 1100으로 뛰어올랐다.

이 문제는 무슨 일이 일어났는지 설명해 준다. 버림 오차는 모두 같은 방향을 향하므로 연산 횟수에 비례하여 꾸준히 쌓인다. 반올림(rounding) 오차는 양쪽 방향을 모두 향하므로 대부분 서로 상쇄된다. 반올림 오차는 연산 횟수의 제곱근에 비례하여 커질 뿐인데, 이것이 바로 중심극한정리(central limit theorem, Section 12.24)가 예측하는 확률보행(random walk)의 행동이다.

시작하기 전에

생각해 보기 어떤 수의 소수 부분 U가 $[0,1)$에서 균등분포(uniform distribution)를 따른다면, 그 수를 버림할 때 오차는 무엇인가? 반올림할 때는 어떠한가? 각각의 평균과 분산(variance)은 무엇인가?

버림은 a를 $\lfloor a \rfloor$로 바꾸며, 이때 오차(참값에서 근삿값을 뺀 값)는 U로서 $[0,1)$에서 균등분포를 따른다. 평균은 $\frac{1}{2}$, 분산은 $\frac{1}{12}$이다. 반올림의 오차는 $(-\frac{1}{2}, \frac{1}{2}]$에서 균등분포를 따르며, 평균은 0, 분산은 $\frac{1}{12}$이다.

차근차근 풀기

1단계 전체 오차

1500개의 수를 더한 뒤의 전체 오차는 1500개의 독립적인 오차의 합이며, 중심극한정리에 의해 근사적으로 정규분포(normal distribution)를 따른다.

- LEP(반올림): 평균 0, 분산 $\frac{1500}{12} = 125$, 표준편차(standard deviation) $\sqrt{125} \approx 11.18$.
- VOZ(버림): 평균 $\frac{1500}{2} = 750$, 표준편차는 똑같이 11.18.

2단계 큰 오차가 생길 확률

> **생각해 보기** LEP의 경우 $E \sim \mathrm{N}(0, 125)$일 때 $\mathrm{P}(|E| > 15)$는 얼마인가? VOZ의 경우 오차는 거의 확실하게 어디에 있는가?

LEP의 경우 $\frac{15}{11.18} \approx 1.342$이므로

$$\mathrm{P}(|E| > 15) \approx 2\big(1 - \Phi(1.342)\big) \approx 2 \times 0.090 = 0.18.$$

VOZ의 경우 오차는 약 750 ± 11이다. 오차의 크기가 15보다 작을 확률은 사실상 영이므로 $\mathrm{P}(|E| > 15) \approx 1$이다. ■

3단계 덧셈을 몇 번 할 수 있는가?

> **생각해 보기** n번 더한 뒤 LEP의 오차의 표준편차는 $\sqrt{n/12}$이다. 언제 $\mathrm{P}(|E| < 10) \geq 0.9$가 성립하는가?

$\mathrm{P}(|Z| < 1.645) = 0.9$이므로 $\frac{10}{\sqrt{n/12}} \geq 1.645$이어야 한다. 즉 $\frac{n}{12} \leq \left(\frac{10}{1.645}\right)^2 \approx 36.95$이므로 $n \leq 443$이다. LEP는 약 443번의 덧셈을 할 수 있다. ■

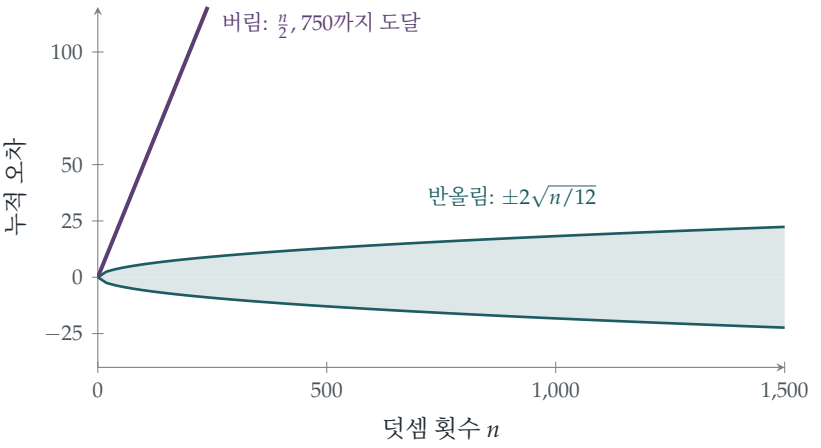


그림 13.6.1 오차가 쌓이는 방식. 버림 오차(자주색)는 모두 부호가 같으며 n에 비례하여 커진다. 1500번 더한 뒤 그 합은 약 750이다. 반올림 오차(청록색 띠, 영을 중심으로 양쪽으로 표준편차의 두 배)는 평균적으로 상쇄되어 $\sqrt{n}$에 비례하여 커질 뿐이다. 1500번 더한 뒤에도 대개 약 ± 22 이내에 있다.

돌아보기

- **각 오차를 모형화한다.** 소수 부분이 균등분포를 따른다는 가정 덕분에 각 오차의 분포가 간단해졌다.
- **평균과 분산을 더한다.** 평균은 더해지며(그 결과 편향 $\frac{n}{2}$가 생긴다), 독립인 오차라면 분산도 더해진다.
- **정규 근사.** 중심극한정리를 이용하여 1500개의 균등분포 오차의 합을 정규분포로 바꾸었다.

숨은 수학

편향과 분산

버림은 **편향**(biased)되어 있다. 즉 버림 오차의 평균은 영이 아니며, 편향은 선형으로 쌓인다. 반올림은 **비편향**(unbiased)이며, 그 오차는 확률보행처럼 행동하여 $\sqrt{n}$에 비례하여 커진다. 연산을 백만 번 하고 나면 그 차이는 500 000 정도의 오차와 300 정도의 오차 사이의 차이가 된다. 이는 실험에서 계통 오차(systematic error, 예를 들어 잘못 보정된 측정 기구)와 무작위 측정 오차(random error)를 구별하는 것과 같다. 측정을 여러 번 하여 평균을 내면 무작위 오차는 줄어들지만 계통 오차는 전혀 줄어들지 않는다.

부동소수점

모든 컴퓨터는 실수를 제한된 정밀도로 저장하며, 모든 산술 연산은 그 결과를 반올림한다. 현대 컴퓨터는 IEEE 754 표준을 따르는데, 이 표준은 표현 가능한 가장 가까운 수로 반올림하고, 두 수의 한가운데인 경우에는 마지막 자릿수가 짝수가 되는 쪽으로 정한다. 바로 이렇게 해야 반올림이 평균적으로 비편향이 되기 때문이다. 그렇더라도 많은 수를 부주의하게 더하면 정확도를 잃을 수 있다. William Kahan의 **보정 합산**(compensated summation) 알고리즘(1965년)은 잃어버린 하위 자릿수를 추적하여 다시 더해 준다. 1991년 걸프 전쟁에서 한 패트리엇 미사일 포대는 날아오는 미사일을 요격하는 데 실패했는데, 그 포대의 시계가 시간을 10분의 1초 단위로 셌기 때문이다. 10분의 1은 이진법으로 정확하게 저장할 수 없는 수이다. 100시간 동안 작동한 뒤 버림 오차는 약 3분의 1초로 커져 있었고, 이는 목표물의 위치를 반 킬로미터 넘게 잘못 판단하기에 충분했다.

밴쿠버 지수

밴쿠버 증권거래소 지수는 1982년 1월 1000.000의 값으로 만들어졌다. 1983년 11월이 되자 시장이 전반적으로 상승했는데도 지수는 약 524까지 떨어져 있었다. 자문가들은 지수가 하루에 수천 번 다시 계산되었고 그때마다 소수점 아래 셋째 자리까지 (반올림이 아니라) 버림되었다는 사실을 알아냈다. 주말 동안 올바르게 다시 계산한 지수는 약 1098에서 다시 시작했다. 이 일화는 수치 계산 강의에서 빠지지 않는 경고 사례이다.

사고 기술

계통 오차와 무작위 오차를 구별하라

부호가 일정한 오차는 선형으로 쌓이고, 평균이 영인 독립적인 오차는 제곱근처럼 쌓인다. 지금 다루는 오차가 어느 종류인지 파악하라.

대학에서 수치해석(numerical analysis)은 계산 과정에서 오차가 어떻게 전파되는지 연구하며, 통계학은 편향과 분산을 구별한다. "편향-분산 트레이드오프(bias–variance trade-off)"는 현대 기계학습(machine learning)의 핵심이다.

더 도전하기

1 **보정된 VOZ.** VOZ가 마지막에 답에 $\frac{\pi}{2}$를 더한다면 그 오차는 어떤 분포를 따르는가? 이제 LEP만큼 좋아졌는가?

2 **소수점 아래 첫째 자리로 반올림하기.** 수를 소수점 아래 첫째 자리로 반올림하고, 버려지는 자릿수가 균등분포를 따른다고 하자. 10 000개의 수를 더할 때 전체 오차의 표준편차를 구하라.

3 **밴쿠버 지수.** 지수를 하루에 3000번 다시 계산하고 한 번 버릴 때마다 평균 0.0005를 잃었다면, 거래일이 250일인 1년 동안 지수는 얼마나 줄어들겠는가?

4 **정확한 분포.** 두 반올림 오차의 합의 정확한 확률밀도함수(삼각형 모양)를 구하라. 세 개의 합에 대해서는 그 개형을 그려 보라. 얼마나 빨리 정규분포처럼 보이기 시작하는가?

더 읽을거리

- **Nicholas Higham, *Accuracy and Stability of Numerical Algorithms*** (SIAM). 반올림 오차에 관한 권위 있는 책으로, 첫 장은 읽기 쉽다.
- **David Goldberg, "What every computer scientist should know about floating-point arithmetic"**, *ACM Computing Surveys* 23 (1991). 고전적인 글로, 온라인에서 무료로 볼 수 있다.
- **Tim Harford, *How to Make the World Add Up*** (Bridge Street). 데이터 속 계통 오차를 다룬다.

13.7 상관관계 만들어 내기

문제

STEP III 2010, 13번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

`step.maths.org/questions/10-s3-q13`

도입

은행은 주식 포트폴리오(portfolio)에서 손실을 볼 가능성이 얼마나 되는지 알고 싶어 한다. 주가는 함께 움직인다. 한 주식이 떨어지면 다른 주식들도 떨어지는 경향이 있다. 가능한 미래를 수천 가지 모의실험(simulation)하려면, 은행은 정확히 알맞은 방식으로 상관된 난수를 만들어 내야 한다. 독립(independent)인 확률변수에서 출발하여, 미리 정한 상관계수(correlation)를 갖는 확률변수를 어떻게 만들어 낼 수 있을까?

이 문제는 세 변수에 대해 그 방법을 한 단계씩 알려 준다. 새로운 변수는 각각 앞서 쓴 독립인 재료들의 결합에, 분산(variance)을 채우기 위한 새로운 독립인 재료를 하나 더 한 것이다. 이 구성은 모습을 감춘 유명한 선형대수의 한 부분, 곧 **숄레스키 분해**(*Cholesky decomposition*)이다. 또한 이 구성은 숨어 있는 제약도 드러낸다. 모든 상관계수의 조합이 가능한 것은 아니다.

출제 범위에 관한 참고

공분산(covariance)과 상관계수는 2019년에 STEP 출제 범위(specification)에서 빠졌다. 공분산은 $\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$이다. 독립인 변수에 대해서는 공분산이 0이고, 공분산은 각 인수에 대해 선형(linear)이다: $\mathrm{Cov}(aU+bV,W) = a\,\mathrm{Cov}(U,W) + b\,\mathrm{Cov}(V,W)$. 또한 $\mathrm{Var}(U) = \mathrm{Cov}(U,U)$이다. 상관계수는 항상 -1과 1 사이에 있다.

시작하기 전에

생각해 보기 Z_1과 Z_2는 독립이다. 둘의 공분산은 얼마이고, 따라서 상관계수는 얼마인가?

독립인 변수는 $\mathrm{E}(Z_1Z_2) = \mathrm{E}(Z_1)\,\mathrm{E}(Z_2)$를 만족한다. 따라서 공분산은 0이고, $\mathrm{Corr}(Z_1, Z_2) = 0$이다.

차근차근 풀기

1단계 두 변수

$\mathrm{E}(Y_2) = 0$이다. Z_1과 Z_2는 분산이 1인 독립인 변수이므로

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2) = \rho_{12}.$$

두 분산이 모두 1이므로 $\mathrm{Corr}(Y_1, Y_2) = \rho_{12}$이다. ■

생각해 보기 Z_1과 Z_2를 서로 수직인 단위벡터(unit vector)로 생각해 보자. Y_2는 어디에 있는가? Y_1과 Y_2 사이의 각의 코사인은 얼마인가?

$Y_2 = \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2$는 $Z_1 = Y_1$과 각 θ를 이루는 단위벡터이고, 여기서 $\cos\theta = \rho_{12}$이다. 상관계수는 변수들을 벡터로 보았을 때 그 사이 각의 코사인이다(그림 13.7.1).

2단계 세 번째 변수

생각해 보기 세 조건 $\mathrm{Var}(Y_3) = 1$, $\mathrm{Cov}(Y_1, Y_3) = \rho_{13}$, $\mathrm{Cov}(Y_2, Y_3) = \rho_{23}$을 a, b, c로 나타내어라.

선형성과 독립성을 이용하면

$$a^2 + b^2 + c^2 = 1, \qquad a = \rho_{13}, \qquad \rho_{12}a + \sqrt{1-\rho_{12}^2}\,b = \rho_{23}.$$

따라서

$$a = \rho_{13}, \qquad b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}, \qquad c = \sqrt{1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2}}.$$ ■

($c \geq 0$이므로 음이 아닌 제곱근을 택한다.)

3단계 임의의 평균과 분산

크기를 바꾸거나 평행이동해도 상관계수는 변하지 않으므로

$$X_i = \mu_i + \sigma_i Y_i \qquad (i = 1, 2, 3)$$

는 평균(mean) μ_i, 분산 σ_i^2, 상관계수 ρ_{ij}를 갖는다. ■

> **잠깐, 직접 해 보기** c에 대한 공식이 성립하려면 제곱근 안의 값이 음이 아니어야 한다. $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$을 넣어 보아라. 무엇이 잘못되는가? 또 세 변수가 서로 모두 강한 음의 상관관계를 가질 수 없다는 것은 왜 이치에 맞는가?

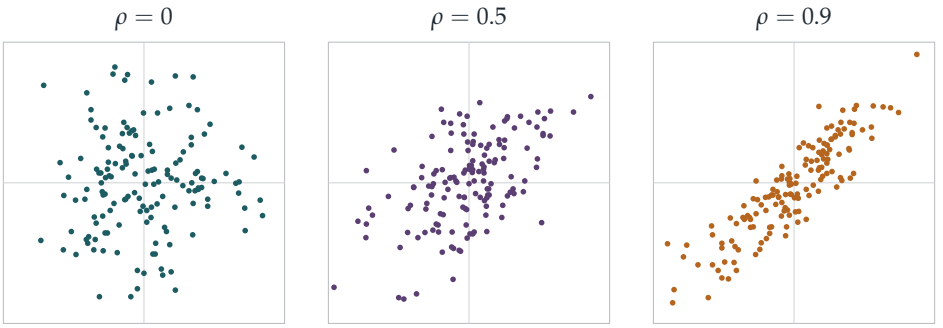


그림 13.7.1 독립인 표준정규(standard normal) 확률변수 Z_1, Z_2로부터 $Y_1 = Z_1$, $Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$라는 방법으로 만든 (Y_1, Y_2)의 표본으로, $\rho = 0$, 0.5, 0.9인 경우이다. 같은 독립인 재료로 우리가 원하는 어떤 상관계수든 만들어 낼 수 있다.

돌아보기

- **한 번에 한 변수씩 쌓아 올린다.** 새 변수는 각각 앞선 재료를 알맞은 양만큼 가져오고, 새로운 독립인 재료를 하나 더했다.
- **공분산은 선형이다.** 모든 조건이 계수에 대한 간단한 방정식이 되었다.
- **상관계수는 코사인이다.** 표준화된 변수를 단위벡터로 생각하면 상관계수가 각이 된다.

숨은 수학

숄레스키 분해

행렬 형태로 쓰면, 이 구성은 **하삼각**(*lower triangular*) 행렬

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

에 대해 $\mathbf{Y} = L\mathbf{Z}$라는 것이고, $\mathbf{Y}$의 상관행렬(correlation matrix)은 $R = LL^{\mathsf{T}}$이다. 대칭행렬을 LL^{T} 꼴로 나타내는 것이 **숄레스키 분해**이다. 이 분해는 프랑스의 군사 기술자 앙드레루이 숄레스키(André-Louis Cholesky)가 1910년 무렵 측지 측량의 방정식을 풀기 위해 찾아냈다(그가 제1차 세계대전에서 사망한 뒤에 발표되었다). 이 분해는 R가 **양의 정부호**(*positive definite*)이면 언제나 가능하며, 그 계산은 우리가 손으로 한 것과 정확히 같다. 각 행을 차례로 풀면 된다. 이것은 금융, 공학, 기후 모델링에서 상관된 변수를 모의실험하는 표준적인 방법이다.

어떤 상관계수가 가능한가?

c의 제곱근 안에 있는 수는 음이 아니어야 한다. 전개하면 그 조건은

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

이고, 이는 상관행렬의 행렬식(determinant)이 음이 아니라는 조건이다. 기하학적으로 보면, 두 개씩 이루는 각이 θ_{ij}(여기서 $\cos\theta_{ij} = \rho_{ij}$)인 세 단위벡터가 실제로 공간 안에 함께 놓일 수 있어야 한다는 뜻이다. $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$이면 각은 모두 약 127°이고, 서로 127°를 이루는 세 벡터는 각의 합이 360°보다 커야 하므로 불가능하다. 일반적으로 어떤 행렬이 상관행렬이 될 수 있으려면 양의 준정부호(positive semi-definite), 즉 모든 고윳값(eigenvalue)이 음이 아니어야 한다(Section 7.2). 불완전한 자료로부터 상관계수를 하나하나 따로 추정하면 이 조건이 깨질 수 있으며, 그러면 통계학자는 가장 가까운 유효한 상관행렬을 찾아야 한다. 이 문제는 2002년에 Nicholas Higham이 해결했다.

골턴, 피어슨, 그리고 상관

프랜시스 골턴(Francis Galton)은 1888년 부모와 자녀의 키가 어떻게 관련되는지를 연구하면서 상관이라는 개념을 도입했고, 칼 피어슨(Karl Pearson)은 1896년 이 문제에서 사용한 적률 공식(product moment formula)을 제시했다. 골턴은 평균으로의 회귀(regression to the mean)도 발견했다(Section 13.1). 피어슨은 1911년 유니버시티 칼리지 런던(University College London)에 세계 최초의 대학 통계학과를 세웠다. "상관관계는 인과관계(causation)를 뜻하지 않는다"라는 경고는 상관이라는 개념 자체만큼이나 오래되었다.

사고 기술

새로운 독립인 재료를 써서 한 단계씩 구성하기

미리 정한 관계를 갖는 대상들을 만들려면 하나씩 추가하되, 각각을 앞서 나온 것들의 결합에 새롭고 독립인 무언가를 더한 것으로 만든다.

이것은 선형대수의 그람–슈미트 과정(Gram–Schmidt process)을 거꾸로 돌린 것이며, 곧 숄레스키 분해이다. 대학에서는 최소제곱법(least squares), 다변량 정규분포(multivariate normal distribution) 이론, 확률과정(random process)의 모의실험에서 나타난다.

더 도전하기

1 **불가능한 상관계수.** 세 확률변수의 두 개씩의 상관계수가 모두 ρ이려면 $\rho \geq -\frac{1}{2}$이어야 함을 보여라. n개의 변수에 대해서는 이에 대응하는 명제가 무엇인가?

2 **행렬식 확인하기.** R가 3×3 상관행렬일 때, $1 - \rho_{13}^2 - b^2 \geq 0$이 $\det R \geq 0$과 동치임을 보여라.

3 **포트폴리오 위험.** 두 주식의 수익률의 표준편차(standard deviation)가 $\sigma_1 = 0.2$, $\sigma_2 = 0.3$이고 상관계수가 ρ이다. 두 주식에 절반씩 투자한 포트폴리오의 표준편차를 $\rho = -1, 0, 1$일 때 각각 구하여라. 분산투자(diversification)는 어떤 효과를 내는가?

4 **재료 되찾기.** 문제에서와 같이 Y_1, Y_2가 주어졌을 때, Z_2를 Y_1과 Y_2로 나타내어라. (이것이 그람–슈미트 단계이다.)

더 읽을거리

- **길버트 스트랭(Gilbert Strang),** ***Introduction to Linear Algebra*** (Wellesley-Cambridge). 양의 정부호 행렬과 숄레스키 분해를 다룬다.
- **Paul Glasserman,** ***Monte Carlo Methods in Financial Engineering*** (Springer). 상관된 변수를 실제로 모의실험하는 방법을 다룬다. 수준이 높지만 앞부분의 장들은 읽을 만하다.
- **Theodore Porter,** ***Karl Pearson: The Scientific Life in a Statistical Age*** (Princeton University Press).

우리가 배운 것

이 장의 첫머리에서 확률(probability)은 모형에서 데이터로 나아가고, 통계학(statistics)은 데이터에서 다시 세계로 돌아간다고 말하였다. 그리고 세 가지 질문을 던졌다. 그것은 실제인가, 아니면 잡음인가? 어떻게 추정해야 하는가? 무작위성은 어떻게 데이터에 들어오는가? 이 장의 일곱 문제가 보여 준 내용은 다음과 같다.

1. 실제인가, 잡음인가?

Section 13.1에서는 네 개의 신문 헤드라인이 숫자를 주장으로 바꾸어 놓았다. 분모 없이 총계를 비교하는 것, 한 해나 하루의 수치에서 추세를 읽어 내는 것, 평균 속에 추세를 감추는 것은 모두 오류였고, 푸아송 모형(Poisson model)은 개수의 자연스러운 변동을 재는 잣대가 되었다. 표준편차(standard deviation)의 8배에 이르는 증가는 거의 확실히 실제였고, 어느 화요일의 사망자 일곱 명은 거의 확실히 그렇지 않았다. 평균으로의 회귀(regression to the mean)와 공정한 비교의 필요성이 이 그림을 완성하였다.

2. 어떻게 추정해야 하는가?

세 문제를 통해 추정 이론을 전개하였다.

- **최대가능도(maximum likelihood)** (Section 13.2): 관측된 데이터가 나올 확률을 가장 크게 하는 미지수의 값을 고른다. 포획-재포획(capture–recapture)에서는 자연스러운 추정값 $\frac{200\times 200}{11} \approx 3636$을 얻었고, 넓게 퍼진 가능도(likelihood)는 이 추정값이 얼마나 불확실한지를 보여 주었다.
- **편향과 분산** (Section 13.3): 잠수함의 수에 대한 두 불편추정량(unbiased estimator)을 살펴보았는데, 하나는 최댓값에, 다른 하나는 평균에 바탕을 둔 것이었다. 충분통계량(sufficient statistic)인 최댓값이 훨씬 더 정밀하였으며, 그 분산(variance)은 $\frac{1}{n}$이 아니라 $\frac{1}{n^2}$처럼 줄어들었다. 같은 추론으로 전쟁 중 독일 전차에 관한 정보기관의 판단을 앞지를 수 있었다.
- **효율과 강건성** (Section 13.4): 균등분포를 따르는 데이터에서는 평균이 중앙값(median)보다 효율(efficiency)이 세 배 높지만, 꼬리가 두꺼운(heavy-tailed) 데이터에서는 중앙값이 이기며, 중앙값은 터무니없는 오차에 대해 강건(robust)하다.

3. 무작위성은 어떻게 데이터에 들어오는가?

무작위성은 의도적으로 더할 수 있다. 확률화 응답(randomised response) (Section 13.5)은 각 사람이 독립적으로 무작위화하기만 하면 응답자 개개인의 사생활을 보호하면서도 모집단을 추정할 수 있게 해 준다. 모두가 동전 하나를 함께 쓴다면 봉우리가 두 개인 혼합분포(mixture)가 생겨 거의 아무것도 알려 주지 못한다. 무작위성은 처리 과정에서도 들어온다. 반올림(rounding) 오차는 서로 상쇄되어 $\sqrt{n}$처럼 커지지만, 밴쿠버 증권거래소(Vancouver Stock Exchange)가 알게 되었듯이 버림(truncation) 오차는 선형으로 누적된다(Section 13.6). 또 무작위성은 설계할 수도 있다. 독립적인 재료들을 숄레스키 구성(Cholesky construction)으로 조합하면 타당한 상관계수의 집합이라면 어떤 것이든 그것을 갖는 변수들을 만들 수 있다(Section 13.7). 다만 모든 상관(correlation)계수의 집합이 타당한 것은 아니다.

4. 통계학 도구 상자

방법	사용한 곳
푸아송 변동	헤드라인 (13.1)
가능도와 그 최댓값	들쥐 (13.2)
불편성과 분산	일련번호 (13.3); 중앙값 (13.4)
충분통계량	일련번호 (13.3)
정규근사(normal approximation)	13.4, 13.5, 13.6
전분산 법칙(law of total variance)	확률화 응답 (13.5)
편향 대 무작위 오차	반올림 (13.6)
공분산의 선형성(linearity of covariance)	상관 (13.7)

다음으로

- 베이즈 정리(Bayes' theorem) (Section 12.2)는 이 장의 방법들에 대한 주된 대안인 베이즈 통계학의 바탕이다.
- 중심극한정리(central limit theorem) (Section 12.24)는 이 장 전체에서 사용한 정규근사를 정당화하며, 코시 분포(Cauchy distribution) (Section 12.22)는 정규근사가 언제 실패하는지를 보여 준다.
- 상관행렬은 양의 준정부호(positive semi-definite) 행렬이며, Section 7.2의 고유벡터(eigenvector)를 이용하여 연구한다.

대학에 가기 전에 더 나아가고 싶다면

- **David Spiegelhalter, *The Art of Statistics***: 일반 독자를 위한 통계적 사고 입문서 가운데 최고이다.
- **Tim Harford, *How to Make the World Add Up***: 뉴스 속 숫자를 읽는 법을 다룬다.
- **John Rice, *Mathematical Statistics and Data Analysis***: 수리통계학의 대학 첫 강좌 교재이다.
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***: 통계학의 큰 아이디어들과 그 기원을 다룬다.

조지 박스(George Box)는 모든 모형은 틀렸지만 그중 일부는 쓸모가 있다고 말하였다. 이 장은 모형을 쓸모 있게 만드는 방법을 보여 주려 하였다. 얼마만큼의 변동을 예상해야 하는지 아는 것, 방법을 여러 번 되풀이했을 때 어떻게 작동하는지로 판단하는 것, 그리고 데이터가 어떻게 생겨났는지 생각하는 것이 그것이다. 이것이 숫자를 지식으로 바꾸는 습관이다.

맺음말

이것으로 이 책은 끝난다. 열세 개의 장에 걸쳐 우리는 STEP 문제를 대학 첫 강좌의 수학으로 들어가는 문으로 삼았다. 정수론과 해석학, 미적분학과 복소수, 대수학과 선형대수학, 미분방정식과 동역학, 조합론, 역학, 확률과 통계가 그것이다. 각 문제는 그 뒤에 무엇이 있는지 묻도록 하는 초대였고, 거의 매번 그 답은 하나의 학문 분야 전체였다. 연필을 손에 들고 여백의 질문을 던지며 끝에 있는 문제를 풀어 보면서 이 절들 가운데 몇 개만이라도 끝까지 공부했다면, 여러분은 이미 수학자처럼 생각하기 시작한 것이다. 나머지 여정은 여러분의 몫이다.

더 도전하기 해답

이것은 완전한 풀이가 아니라 간단한 답과 힌트이다. 결과와 핵심 단계를 제시하므로 자신의 풀이를 확인하는 데 쓸 수 있다. 답을 보기 전에 각 연습문제를 제대로 풀어 보아라.

13.1 신문 머리기사 읽기

1. $1 - \sum_{k=0}^{6} \mathrm{e}^{-2.96} \frac{2.96^k}{k!} \approx 0.0315$이므로, 사망자가 일곱 명 이상인 날은 1년에 약 11일이다. $\mathrm{P}(X \geq 9) \approx 0.0035$이므로 이런 날은 1년에 약 하루이다.

2. 사망자 수는 $150K$이고, K는 평균(mean)이 1.2인 포아송(Poisson) 분포를 따른다. 평균은 180, 분산(variance)은 $150^2 \times 1.2 = 27\,000$, 표준편차(standard deviation)는 약 164이다. 사망자가 125명 이하인 해는 추락 사고가 한 번도 없는 해이고, 그 확률은 $e^{-1.2} \approx 0.30$이므로 전혀 이상할 것이 없다. 그러나 이 모형에서 사망자가 정확히 125명인 것은 불가능하다. 사망자는 150의 배수로만 생기기 때문이다. 평균과 표준편차로는 이 사실이 드러나지 않는다. 군집(clustering) 때문에 분포가 울퉁불퉁해져서, 모든 확률이 0, 150, 300 등에 몰려 있다.

3. $z = \frac{1349.5 - 1080}{33} \approx 8.2$이므로 확률은 10^{-16} 정도이다. 일정한 비율로는 1350을 설명할 수 없다. 무언가가 실제로 바뀌었다(또는 사망자를 세는 방식이 바뀌었다).

4. 예를 들면 다음과 같다. (i) "총계만으로는 비행기와 걷기 중 어느 쪽이 더 안전한지 알 수 없다"; (ii) "1990년 항공 사고 사망자, 정상 범위 안"; (iii) "화요일 보행자 사망 7명: 나쁜 날이지만 1년에 약 열한 번 있는 일"; (iv) "1990년 보행자 사망 급증: 원인 조사 필요".

13.2 들쥐 세기

1. $\frac{p_N}{p_{N-1}} = \frac{(N-M)(N-C)}{N(N-M-C+R)}$이고, 이 값은 $NR \leq MC$일 때 그리고 그때에만 1 이상이다. 따라서 p_N은 $N = \frac{MC}{R}$까지 증가하다가 그 뒤로 감소하며, 최댓값은 이 값의 정수 부분에서 나온다.

2. a명의 여자와 b명의 남자 중에서 c명으로 이루어진 위원회의 수를 여자의 수 j에 따라 나누어 센다. 각 j에 대하여 $\binom{a}{j}\binom{b}{c-j}$가지이다.

3. $\frac{\mathrm{d}}{\mathrm{d}p}\big(h \ln p + (n-h)\ln(1-p)\big) = \frac{h}{p} - \frac{n-h}{1-p} = 0$에서 $p = \frac{h}{n}$을 얻는다.

4. $R = 0$이면 p_N은 끝없이 증가하고, 추정값(estimate)은 무한대가 된다. 이 방법은 두

번째 표본에 표지된 동물이 충분히 들어 있을 때에만 통한다. 다시 잡힌 개체가 적으면 추정값이 매우 불안정하다.

13.3 독일 전차 문제

1. 최댓값을 M이라 하면 $\mathrm{E}(M) = \frac{nN}{n+1}$, $\mathrm{Var}(M) = \frac{nN^2}{(n+1)^2(n+2)}$이므로 $\mathrm{E}(M-N)^2 = \frac{2N^2}{(n+1)(n+2)}$이다. 이것은 $\mathrm{Var}(Z_1) = \frac{N^2}{n(n+2)}$의 $\frac{2n}{n+1}$배이므로 더 나쁘다. 가장 좋은 배수는 $c = \frac{\mathrm{NE}(M)}{\mathrm{E}(M^2)} = \frac{n+2}{n+1}$이고, 이때 평균제곱오차(mean squared error)는 $\frac{N^2}{(n+1)^2}$로 Z_1보다 약간 더 좋다.

2. $\mathrm{E}(\min) = \frac{N}{n+1}$이고 min의 분산은 M의 분산과 같으므로 $\mathrm{Var}\big((n+1)\min\big) = \frac{nN^2}{n+2}$이다. 이 값은 n이 커져도 줄어들지 않으므로, 좋지 않은 추정량(estimator)이다.

3. 최댓값이 m 이하인 것은 k개의 수가 모두 $\{1,\dots,m\}$에서 나올 때 그리고 그때에만이다. $k=1$이면 추정량은 $2X-1$이고, 그 평균은 $2\cdot\frac{N+1}{2}-1 = N$이다.

4. $98 + \frac{98}{5} - 1 = 116.6$이므로 탱크는 약 117대이다.

13.4 평균인가, 중앙값인가?

1. 어느 값이 중앙값(median)인지 고르고($2n+1$가지), 나머지 중 어느 n개가 그보다 작은지 고른다($\binom{2n}{n}$가지). 따라서 확률은 $(2n+1)\binom{2n}{n}y^n(1-y)^n\,\delta y = \frac{(2n+1)!}{(n!)^2}y^n(1-y)^n\,\delta y$이다.

2. $\mathrm{Var}(Y) = \frac{1}{4(2n+3)} = \frac{1}{20}$이고 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)} = \frac{1}{36}$이다. 평균이 더 좋다.

3. $\mathrm{Var}(\max) = \mathrm{Var}(\min) = \frac{m}{(m+1)^2(m+2)}$이고 $\mathrm{Cov}(\max,\min) = \frac{1}{(m+1)^2(m+2)}$이므로 $\frac{1}{2(m+1)(m+2)}$를 얻는다. 이 값은 $\frac{1}{m^2}$처럼 줄어들어, 평균의 $\frac{1}{12m}$보다 훨씬 빠르게 작아진다. 균등분포(uniform distribution)를 따르는 자료에서는 탱크 문제에서처럼 양 끝값이 정보를 담고 있다.

4. 평균은 약 18.6으로, 오타 하나 때문에 망가진다. 중앙값은 10.1로, 영향을 받지 않는다.

13.5 곤란한 질문 던지기

1. 참인 답을 할 확률을 $t = ap + bq$라 하자. "예"라고 답할 확률은 $\frac{1}{2}t + \frac{1}{2}(1-t) = \frac{1}{2}$이므로 $\mathrm{P}(\text{앞면} \mid \text{예}) = t$이고, 이것은 $2t=1$일 때 그리고 그때에만 $\frac{1}{2}$이다. 같은 쌍둥이가 두 번 답하면, "예" 다음에 "아니요"라고 답할 확률은 동전이 앞면이든 뒷면이든 $pa(1-a)+qb(1-b)$이므로 사후확률(posterior probability)은 $\frac{1}{2}$이다. 두 번 모두 "예"라고 답할 확률은 pa^2+qb^2(앞면)와 $p(1-a)^2+q(1-b)^2$(뒷면)이고, 이 둘은 $2(ap+bq)=1$일 때 그리고 그때에만 같다.

2. $\mathrm{P}(\text{예}) = \frac{3}{4}\pi + \frac{1}{4}(1-\pi)$이므로 $\hat{\pi} = \frac{0.4-0.25}{0.5} = 0.3$이다. 표준오차(standard error)는 $\frac{\sqrt{0.4\times 0.6/400}}{0.5} \approx 0.049$이다.

3. $\mathrm{E}(Y^2) = \sum_c \mathrm{P}(C=c)\big(\mathrm{Var}(Y \mid c) + \mathrm{E}(Y \mid c)^2\big)$이다. 여기서 $\mathrm{E}(Y)^2 = \big(\sum_c \mathrm{P}(C=c)\mathrm{E}(Y \mid c)\big)^2$을 빼면 $\mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$를 얻는다.

4. Z_2는 각각 확률 $\frac{1}{2}$로 $\frac{n}{6}$(앞면) 또는 $\frac{5n}{6}$(뒷면)에 가깝다. 따라서 Z_2가 $\frac{n}{2}$ 근처에 있는 일은 거의 없는데도 $\mathrm{P}(Z_2 > \frac{n}{2}) \to \frac{1}{2}$이다.

13.6 쌓이는 오차

1. 각 버림 오차(truncation error)는 $(-1, 0]$에서 균등분포를 따른다. 각각에 $\frac{1}{2}$를 더하면 $(-\frac{1}{2}, \frac{1}{2}]$에서 균등분포를 따르게 되어, 반올림(rounding)과 똑같아진다. 따라서 보정한 VOZ는 LEP만큼 좋다.

2. 각 오차는 $(-0.05, 0.05)$에서 균등분포를 따르고 분산은 $\frac{0.1^2}{12}$이므로, 합의 표준편차는 $\sqrt{10\,000 \times 0.01/12} \approx 2.9$이다.

3. $3000 \times 250 \times 0.0005 = 375$점이다.

4. 오차가 두 개이면 $[-1, 1]$에서 $1 - |x|$이다. 세 개이면 $|x| \le \frac{1}{2}$에서 $\frac{3}{4} - x^2$이고 $\frac{1}{2} \le |x| \le \frac{3}{2}$에서 $\frac{1}{2}\big(\frac{3}{2} - |x|\big)^2$이며, 이미 정규분포(normal distribution) 곡선에 가깝다.

13.7 상관관계 만들어 내기

1. 표준화된(standardised) 변수에 대하여 $\mathrm{Var}(X_1 + X_2 + X_3) = 3 + 6\rho \ge 0$이므로 $\rho \ge -\frac{1}{2}$이다. n개의 변수에 대해서는 $n + n(n-1)\rho \ge 0$에서 $\rho \ge -\frac{1}{n-1}$을 얻는다.

2. $b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}$라 하면 $(1-\rho_{12}^2)(1-\rho_{13}^2-b^2) = 1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} = \det R$이다. $1 - \rho_{12}^2 > 0$이므로 두 조건은 동치이다.

3. $\mathrm{Var} = \frac{1}{4}(0.04 + 0.09 + 0.12\rho)$이므로, $\rho = -1, 0, 1$일 때 표준편차는 각각 0.05, 0.18, 0.25이다. 두 주식이 완전한 상관관계(perfectly correlated)에 있지 않은 한, 두 주식을 섞어 보유하는 것은 두 위험의 평균(0.25)보다 덜 위험하다.

4. $Z_2 = \dfrac{Y_2 - \rho_{12}Y_1}{\sqrt{1-\rho_{12}^2}}$.

Beyond STEP

STEP은 대부분의 학생이 학창 시절에 치르는 수학 시험 가운데 가장 어려운 시험이다. 또한 가장 훌륭한 시험 가운데 하나이기도 하다. 문제는 짧고 낯설며 깊이가 있고, 그중 가장 뛰어난 문제들은 대학 수학의 작은 조각이 모습을 감추고 있는 것이다.

이 책은 1987년 첫 시험에 앞서 나온 예시 시험지(specimen papers)부터 2025년 시험지까지 가운데 가장 흥미로운 STEP 문제들을 골라, 각 문제를 하나의 문으로 삼는다. 모든 문제는 뛰어난 학생이 풀어 가면서 던질 법한 질문들과 함께 꼼꼼하게 풀이하고, 그런 다음 책은 그 문을 지나 한 걸음 더 나아간다. 그 아이디어가 어디에서 왔는지, 대학에서는 무엇으로 발전하는지, 그리고 다음에는 무엇을 읽으면 좋은지를 살펴본다.

이 책은 STEP을 준비하면서 기출 문제 연습 이상의 것을 원하는 재능 있는 예비 수학자를 위해 쓰였다. 수학과 학부생처럼 생각하기 시작하고, 앞으로 무엇이 자신을 기다리고 있는지 알아보고 싶은 학생들을 위한 책이다.

- 순수수학(pure mathematics), 역학(mechanics), 확률(probability)과 통계(statistics)를 아우르는 13개의 장
- 이 36개의 STEP 기출 문제와 각 문제의 안내식 풀이
- 모든 문제 뒤에 숨은 대학 수학
- 답이 실린 후속 연습 문제와 더 읽을거리

전체 13개 장 가운데 제13장. 책 전체와 모든 장,
모든 풀이를 다음 주소에서 무료로 읽고, 내려받고, 공유할 수 있다.

beyondstep.keithmcnulty.org

본문은 퍼블릭 도메인(CC0 1.0)으로 공개한다. STEP 문제 © Cambridge University Press & Assessment.

